ESSAYS ON RECOVERING LATENT CHARACTERISTICS USING NONPARAMETRIC AND SEMIPARAMETRIC METHODS

by

Ryan Bush

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

May, 2016

Abstract

This dissertation explores applying nonparametric and semiparametric methods to recover latent characteristics in various settings. The first chapter studies an auction market where latent effort is selected by the bidders. Recently, states have experimented with offering nonstandard contracts through procurement auctions in an attempt to reduce long-term costs on projects. With no framework in place to study the effects of these of changes, these innovative projects lack the rigorous analysis needed to see widespread implementation. The effects on effort, immediate costs, and potential future costs are all relevant to policy debates about the use of nonstandard contracts. This paper formally addresses the use of nonstandard warranty contracts in procurement auctions by building latent effort choice into an auction model. Taking my model to data from Michigan, I find that warranty auctions encourage roughly 30% of the lowest cost firms to undertake higher effort work in order to reduce future expected costs from repair. Although implementing these contracts across all auctions would have cost an additional $68.7 million in immediate costs over my sample period, Michigan would have saved roughly $3 million annually in the long-run due to the elimination of future repair work by the state.

The second chapter employs panel data to recover latent behavior related to the health and drug use of survey respondents. The misreporting problem of drug use in self-reported surveys can severely affect the validity of estimation results in empirical work. In this paper we use an eigen-decomposition method to nonparametrically estimate the misclassification errors under various assumptions and settings. We use the longitudinal data of NLSY97 and focus on the

years from 2005 to 2009, when the cohort is aged in their mid-20s. We find that the overall proportion of participants who actually use marijuana is higher than the reported proportion. Moreover, participants inclination to misreport their drug use status is related to their current and previous actual drug use as well as their habit for misreporting in surveys. In general, males are more likely to underreport their drug use than females.

The third chapter builds on the recent developments of nonparametric methods for measuring asset price volatility that allow researchers to distinguish between daily volatility arising from discrete price changes (jumps) and that arising from continuous price changes. We investigate the asymmetric effects of signed jumps and semivariance on future volatility in the US Treasury futures market. We compare and contrast our findings to the empirical work in Patton and Sheppard (2015) which investigated the equities market. Additionally, we are able to comment on changes in the asymmetric effects as one moves along the yield curve by employing 5-, 10-, and 30-year Treasury futures data. Past work has shown important distinctions between the foreign exchange, equities, and Treasuries market with respect to other aspects of realized volatility forecasting, but the literature has not considered the differences when examining semivariance. We find that negative semivariance has much stronger effects on future volatility than positive semivariance for longer dated maturities. Additionally, we find that both positive and negative jumps dampen future volatility, although to different degrees.

Keywords: Auctions, Nonparametric Methods, Measurement Error, Forecast

JEL classification: C14, C53, C57, E47, I1

Advisors: Jorge Balat

Richard Spady

Jonathan Wright

Acknowledgements

I am deeply indebted to my advising committee, Jorge Balat, Richard Spady, and Jonathan Wright, for their support and guidance. They have shaped the way I view economics and the world, although I will not hold it against them if they try to distance themselves from that claim. Their thoughtfulness and generosity made graduate life more pleasant, and I hope they reflect back on their time with me as fondly as I do my time with them.

The second chapter of this dissertation is joint work with Yingyao Hu and Yajing Jiang, whom I thank for their collaboration. Special thanks are due to Robert Barbera and Jon Faust, who each taught me important lessons about the consequences and limitations of economic policy. Robert Moffitt has shown tremendous care and concern for all graduate students. The time and energy he invested into my well-being and career have enabled me to pursue a life that I find rewarding. Additionally, I have benefited from discussions on research and life over the course of my time at Johns Hopkins with Christopher Carroll, Ying Chen, Daniel Garcia, Burçin Kısacıkoğlu, Elena Krasnokutskaya, Sohini Mahapatra, Chris Martin, Gwyn Pauley, and Matt Wiswall.

I am especially grateful to my loving family and friends for their unwavering support. I would not have been able to finish this dissertation without their kindness, patience, and encouragement.

Table of Contents

List of Tables

List of Figures

Chapter 1

Latent Effort in Procurement Auctions with Items Under Warranty

1.1 Introduction

State governments have begun experimenting with the use of warranties in procurement auctions, leaving the winning contractor responsible for the project maintenance after initial completion.[1] Although warranties are understood in the posted-price market, their effects on outcomes and competition in auctions have been ignored (Lutz and Padmanabhan (1998)). Despite the lack of academic analysis, these contracts continue to receive considerable attention in the real world. The Michigan Department of Transportation (MDOT), for example, spends over $200 million annually at procurement auctions where the job being procured comes with a warranty for the work (roughly 32% of the dollar value of all their contracts). Ohio, Florida, and South Carolina are other early adopters, each spending over $100 million on Department of Transportation

[1] Throughout this chapter, I use 'procurement auction' to refer to an auction where contractors submit bids to complete a task for the state, and the lowest bidder wins the right to complete the task.

warranty contracts annually (Bayraktar, Cui, Hastak, and Minkarah (2004)).

The main goal in implementing warranties through procurement auctions is to reduce states' long-term costs and transfer risk from the state to the contractor. How does the introduction of these warranties change the way contractors bid, and should others consider using them? I develop a structural model and use data from Michigan procurement auctions in order to answer these questions as well as analyze the long-term benefits of implementing such auctions more broadly.

A key contribution of my paper is to study how a contractor's effort plays a role in procurement auctions, as is the case with auctions where a warranty is assigned to the project. Unfortunately, a contractor's effort is not directly observable by the state or econometrician, complicating attempts to model and estimate it. To better understand the effects of warranties on contractors' effort and bids, as well as the benefits to the state, it is essential to consider the incentives faced by both parties involved in the transaction. In the traditional setting where there are no warranties, the state is responsible for maintaining and correcting any issues related to the project that arise after completion. Even when construction meets the standards set out by the state, repairs and upkeep are common, according to engineers overseeing these projects. These future repairs are a major burden for the state. Under the setting where the project has a warranty, though, the burden of project maintenance is transferred to the winning contractor, relieving the state of any maintenance costs over the life of the warranty.

Conversely, the contractor is incentivized to complete the task using the least expensive methods given the requirements, since any price markup of the bid

over the contractor's cost is profit to the contractor. In standard auctions, least expensive simply means using the level of effort that is required to fulfill the minimum standards of the job. In other words, contractors are not rewarded for producing work that goes beyond the requirements. When a warranty is attached to the job, "least expensive" takes on a more complex interpretation. From the contractor's perspective, undertaking higher effort is costly in the short run; it requires the contractor to use expensive technologies and techniques. This higher level of effort leads to a lower probability that a contractor will need to undertake corrective action in the future. This tradeoff between increases in immediate cost and lower expected future costs, both due to higher effort, defines the tension a contractor faces when determining the optimal level of effort. Specifically, I model immediate and expected future costs, making both types of costs functions of a contractor's endogenous effort choice. This cost structure introduces the tradeoff in effort choice necessary for studying these auctions that is not present in the current literature.

In this first attempt to empirically study the endogenous choice of effort by the contractor, I directly compare the results to standard auction contracts in order to inform policymakers as to the implications. By nesting the standard auction model as a special case, I study how warranties affect outcomes relative to the standard auction setting, as well as how a misspecified model where effort was not optimally chosen (standard auction model) leads to results that inform policymakers to undertake actions against their best interests. By keeping the model as simple as possible while accounting for the institutional details associated with highway procurement, my model facilitates a transparent interpretation of the results.

In order to estimate the model with latent effort choice in auctions, I use a unique data set containing ex post data that captures outcomes related to the effort choices of the contractors. This data set, provided by MDOT, lists all projects that were procured with a warranty as well as measures of ex post corrective action taken by the contractors. This corrective action information is essential to empirically estimating a model with latent effort choice by the contractor. My identification strategy relies on the quasi-natural experimental setting that brought about the introduction of warranties. By estimating the underlying latent cost distribution of contractors and using information related to ex post corrective action, I am able to uncover the proportion of jobs associated with extra effort, the immediate cost increase due to the effort, and the potential long-term effects of adopting better technology and techniques.

I find that the introduction of warranties leads to higher effort work. Overall, these findings support general theoretical predictions about how warranties affect effort and serve as evidence that the model of endogenous effort choice is an important next step in understanding the mechanism currently being used. In total, roughly 30% of firms are willing to optimally undertake higher effort to offset future costs associated with corrective action.

Counterfactual analysis shows that in order to universally implement these contracts, it would have increased the immediate costs to Michigan by $68.7 million over the period I study ($17.2 million annually). Over the longer term, though, Michigan would actually realize savings from using warranty auctions, which accounts for the fact that MDOT would no longer be responsible for the maintenance of the projects over the life of the warranty. Under some additional assumptions, I show that the net benefit when accounting for reduced future

costs by the state, related to managing corrective action in the standard setup, exceeds $3 million a year. Additionally, I find potential long-term benefits that could arise from widespread adoption of these contracts leading to cost reductions when performing high effort work.[2] If the additional immediate cost markup from undertaking high effort work were to fall by 10%, total costs from employing warranty auctions would fall an additional $6.1 million, further encouraging the use of these innovative contracts.

Highway procurement auctions, with their rich set of project-level observable covariates, are a natural choice to study the effects of policy changes on auction outcomes, and there is good reason to believe these results would carry over to other settings. Bidders in many other settings, such as telecommunications, energy, and commercial construction, face similar demands from the state in terms of performance and effort. Additionally, the bidders in these alternative settings have some commonalities with highway contractors in terms of fixed and variable costs, making the model I employ directly relevant.

The structure of my paper proceeds as follows. I briefly discuss the existing literature in Section 2. Next, I present my model of bidding in warranty auctions in Section 3. Section 4 discusses the data I use from MDOT, highlighting the novel data I introduce to recover information related to latent effort choice. Section 5 outlines my empirical strategy and presents the results from my estimation. I undertake two counterfactual experiments to better understand the costs and benefits to the state in Section 6. Section 7 concludes.

[2]These cost reductions are related to the R&D induced by the warranty auctions and the learning-by-doing aspect of undertaking high effort work. See Tiererova (2013) for a formal model of learning-by-doing in procurement auctions.

1.2 Brief Literature Review

My paper brings together strands of literature in empirical auctions while adding to the emerging literature on warranties. First, I discuss recent papers on non-standard auction contracts, as they serve as the primary motivation for this paper. Next, I relate my work to other studies on highway auctions and briefly discuss the methodological foundations of my approach. I conclude with a brief description of the highway engineering literature, which provides evidence that both motivates and supports my modeling decisions.

Closely related to the subject of this paper, there has been increased attention devoted to what are called "innovative contracts." Whereas the standard highway procurement contract is awarded to the lowest qualified bidder for a set date of completion, at which point they are relieved of all of their responsibilities to the state, innovative contracts have recently been employed to more properly align the incentives of the contractors and state by adding caveats to contracts.

In Lewis and Bajari (2011a), the authors considered scoring auctions, where the bidder submits both a dollar value and a time of completion. The data showed that this selection mechanism, known as an A+B contract, substantially reduced the time contractors took to complete their jobs when compared to the standard auction. Similarly, Lewis and Bajari (2011b) evaluated how time incentives in contracts "motivate adaptation." Their work quantified these effects by applying incentives or disincentives in ways that help states accomplish their goal of time completion.

The literature on warranties and producer liability is saturated with theoretical results but is still less developed on the empirical side. Two of the major theoretical papers include Spence (1977) and Cooper and Ross (1985), the latter of which considers effort in the context of upkeep by a purchaser of a good in a posted-price market. Of the limited empirical work related to warranties, there is one recent paper that is especially of interest to my study. Roberts (2011) looked at how warranties can substitute for reputation in online auctions. He did not fully develop an auction model, rather he looked at some reduced form results related to the policy switch from the auction platform not having any warranties to having all warranties. In this way, his paper is related to Lewis (2011) in that both papers tried to understand what observable features affected the selling price, but neither of these papers went so far as to estimate the primitives of a full auction model. Roberts (2011) found that warranties did not reduce the premium commanded by sellers with high reputations, except for those sellers with the highest reputations.

Some early work on highway auctions focused on trying to detect collusion and bid rigging. Porter and Zona (1993) laid the foundation for empirical work on highway auctions by considering data from procurement auctions in New York. Bajari and Ye (2003) estimated models of both competition and collusion to try to detect if the observable bidders were colluding. An interesting feature of their work was the incorporation of information from industry specialists, as I also use conversations with employees at MDOT and private contracting firms to motivate many of my modeling decisions.

More recently, a few papers have captured more nuanced aspects of the procurement process, including auction dynamics (see Jofre-Bonet and Pesendorfer

(2003) and Balat (2012)). Bajari, Houghton, and Tadelis (2006) approached the dynamic cost considerations in a similar way to my paper. When bidders submit their bids in highway auctions, they do so by line item (groundwork, paving, trenching, etc.) even though the total of all line items is used to determine the winner. Their paper considered a framework where contractors were able to anticipate changes to the quantity of inputs demanded by the state for each line item and strategically bid such that they would be optimally compensated at a future date for these changes. My model treats the expectation of future corrective action costs as being known to the contractors, and the contractors alter their bids accordingly to account for such expenses.

I use the nonparametric techniques first developed by Guerre, Perrigne, and Vuong (2000), of which many papers extended the results to analyze alternative paradigms (see Li, Perrigne, and Vuong (2002), Campo, Perrigne, and Vuong (2003), and Li, Perrigne, and Vuong (2000)). I develop an identification and estimation strategy that exploits having data on outcomes under different auction structures, in my case warranty and standard auctions. This reliance on different structures is similar to Lu and Perrigne (2008), which used outcomes in timber auctions for both English and sealed-bid auctions to learn about bidders' costs and risk aversion.

Although warranties on highways have not been studied by economists, these innovative contracts have been examined by many engineers. Cui, Johnson, Sharma, and Bayraktar (2010), Thompson, Anderson, Russell, and Hanna (2002), Gharaibeh and Shirazi (2009), and Bayraktar, Cui, Hastak, and Minkarah (2004) all looked at how states implemented these new warranty contracts. These papers discussed the differences between warranty contracts and standard

contracts, emphasizing that warranty contracts allowed the contractor some freedom with her materials. The papers also tended to agree on the reasons for such contracts: higher quality, lower life-cycle costs, and a redistribution of risk. Some of the papers discussed methods for modeling the risk associated with the warranties while others discussed survey results from contractors and state agencies. One paper used those surveys to estimate that warranties account for between a 0 to 15% increase in bids. Overall, these papers provide a helpful technical perspective when trying to understand new warranty contracts.

1.3 Model

I introduce latent effort into an auction model framework to account for how effort affects total project costs. I draw comparisons between the auction model where the project being procured comes with a warranty (henceforth called "warranty auctions") and its standard auction counterpart, and show the standard model is nested as a special case. This new model provides the flexibility needed to analyze my setting where the level of effort is endogenously selected by the bidder. It is natural to first discuss the standard auction model in order to have a grasp on what information is available in that environment as well as provide the basic intuition for the more complicated model needed to consider warranties.

To be as clear and explicit as possible, I adopt notation common to the auction and econometrics literature. Random variables in my model, the bids and cost draws, are denoted by capital letters whereas their realizations are in lower

case. When dealing with distributions of variables observed by the econometrician, I use $G(\cdot)$ for the CDF and $g(\cdot)$ for the pdf, whereas distributions that are unobserved are denoted by $F(\cdot)$ and $f(\cdot)$. This convention will be convenient when walking through my estimation strategy that involves taking observed bids and backing out the unobserved costs. Lastly, I use bold to distinguish vectors from scalars.

I start my analysis of the standard procurement auction by considering the expected profit function for player i. For simplicity, I assume there is just one auction with $n \geq 2$ bidders. I will introduce additional notation in the estimation section to account for multiple auctions with different numbers of bidders across auctions, but for now I have:

$$E[\pi_i|c_{1i}] = (b_i - c_{1i})Pr(b_i < b_{-i})$$

where $-i$ represents all of bidder i's competitors. In the equation above, b_i is bidder i's submitted bid, c_{1i} is bidder i's cost draw which is assumed to come from the distribution $F_{c_1}()$ with compact, convex support $[\underline{c}, \overline{c}]$.[3] The term $Pr(b_i < b_{-i})$ represents the probability that bidder i's bid is less than her competitors' bids, thus winning the auction in the procurement setting. I assume cost draws are i.i.d. with bidders all drawing from the same distribution. This puts my framework under the umbrella of the symmetric independent private values (IPV) paradigm.[4]

Using $\beta()$ to denote the symmetric Bayesian Nash equilibrium bid function,

[3] I have a 1 in the subscript of the cost draw because when I expand my model to the case with warranties, I include an additional (future) cost draw, c_{2i}, and I want to make clear which cost draws are the same across auction formats and which cost draws are new.

[4] See Paarsch and Hong (2006) for details related to this framework.

I take first order conditions with respect to b_i:

$$\frac{\partial \pi}{\partial b_i} = [1 - F_{c_1}(\beta^{-1}(b_i))]^{n-1} + (b_i - c_{1i}) \frac{\partial [1 - F_{c_1}(\beta^{-1}(b_i))]^{n-1}}{\partial b_i} = 0 \qquad (1.1)$$

with the following solution (see Paarsch and Hong (2006)):

$$\beta(c_{1i}) = c_{1i} + \frac{\int_{c_{1i}}^{\bar{c}} [1 - F_{c_1}(x)]^{n-1} dx}{[1 - F_{c_1}(c_{1i})]^{n-1}}$$

Although standard auctions were the only mechanism employed by departments of transportation (DOT) leading up to the 1980s, some states started experimenting with alternative mechanism in order to provide incentives for contractors to change behavior by shifting risk from the state to the contractors. North Carolina was the first state to have their DOT use warranty auctions for highway pavement marking in 1987. By 1999, over 20 states had experimented with using warranty auctions in various ways. Michigan is regarded as one of the most active users of warranty auctions, which is one of the main reason for using its data in my empirical analysis (Bayraktar, Cui, Hastak, and Minkarah (2004)). Minnesota, another leading state in terms of warranty auction implementation, described the use of warranty auctions as follows: "The goal of instituting warranties on projects is to improve the quality and durability of products by allowing a longer time frame to accept work constructed by contractors. This process is also aimed to encourage contractors to improve their construction techniques and use better equipment to meet the warranty requirements (Construction and Contracting (2006))."

The desired increase in effort and increased use of innovative techniques that motivate the employment of warranty auctions seem to play out in practice.

When interviewed, contractors commented on different ways they changed their behavior when constructing a warranty job, including getting more design input, using better equipment, bringing innovative technology and methods to the job, as well as general quality conscious construction (Bayraktar, Cui, Hastak, and Minkarah (2004)). When the procured job involves paving, DOTs found that warranty contracts induce contractors to take preemptive action to improve poor underlying soil conditions (Management (2007)). Contractors have also found long-term benefits from investing in new formulations of concrete, which include silica fume, steel, and organic fibers, in order to raise the strength of the road for longevity (Skinner Jr (2008)).[5]

These increases in effort come with consequences: the near term increases in the costs, and thus bids, of contractors. All responses from contractors in a survey study indicate they bid just as much or more in the warranty setting than in the standard auction setting (Bayraktar, Cui, Hastak, and Minkarah (2004)).

Motivated by these findings, I model these contracts by adding flexibility to the cost function of each firm in order to account for both the potential immediate increases in costs as well as potential future costs of corrective action.[6] More specifically, I allow firms to choose to undertake extra effort today, which could represent using more innovative and costly materials, employing more specialists, or other costly management decisions, in order to reduce their probability of having to correct for issues that fall under the warranties.

[5]While high-performance concrete has been measured to have strength of 10,000 p.s.i, the new formulations have managed to triple that strength and offer far superior durability.

[6]Corrective action is the official term used by MDOT to describe any required repairs the contractor must undertake during the lifetime of the warranty contract.

More formally, I model firms as having a choice of employing extra effort, $\bar{e}$, that directly scales their immediate costs in the following linear fashion: $(1+\bar{e})c_{1i}$. A contractor can also choose to undertake no extra effort to complete the project, in which case her immediate cost would simply be c_{1i}. To restate the choice in notation, contractor i has a choice of exerting extra effort, e_i, where $e_i \in \{0, \bar{e}\}$.

Next, I model the future cost component of having to undertake corrective action as $ProbCA(e_i = 0)\delta c_{2i}$, for the case when no extra effort is undertaken. Here, δ is the proportion of the job needing corrective action, $ProbCA(e_i = 0)$ is the probability of corrective action in the future given the choice of not undertaking any extra effort today, and c_{2i} is a second cost draw made in the future if one needs to undertake corrective action. I model the second cost draw as coming from the same distribution as the original cost draw, $F_{c_2}() = F_{c_1}()$. Additionally, I assume that the two cost draws are independent. If one thinks about these cost draws as being some idiosyncratic costs related to the bidder at the time of the original auction, it is reasonable to assume there will not be dependence between this draw and the one they receive years later when they need to undertake corrective action. Conversations with industry specialists give credibility to this assumption. Contractors use extra effort $\bar{e}$ to eliminate their chance of having to undertake corrective action in the future (i.e. $ProbCA(e_i = \bar{e}) = 0$), and thus there is no expected future cost associated with that decision. This assumption facilitates a clear way of interpreting how much extra effort is actually captured by $\bar{e}$ and what the tradeoffs are in this case. Thus the expected profit function in the warranty auctions is:

$$E[\pi_i|c_{1i}] = \begin{cases} (b_i - c_{1i} - ProbCA(e_i = 0)\delta\mu_c)Pr(b_i < b_{-i}) & \text{if player } i \text{ chooses to not exert extra effort } (e_i = 0) \\ (b_i - (1+\bar{e})c_{1i})Pr(b_i < b_{-i}) & \text{if player } i \text{ chooses to exert extra effort } (e_i = \bar{e}) \end{cases}$$

A player's expected cost when she chooses no extra effort is $c_{1i} + ProbCA(e_i = 0)\delta\mu_c$ and with extra effort is $(1+\bar{e})c_{1i}$.[7]

The player can solve the maximization with respect to e_i while abstracting away from the choice of b_i, because whatever effort level she may want to employ when formulating her bid is the same effort level she will employ if she is the winner. No new private information arrives between the formulation of bid and initial construction by the winner, therefore there will be internal consistency. A player will choose her optimal effort e^* following the rule:

$$e^* = \begin{cases} 0 & \text{if } (1+\bar{e})c_{1i} - c_{1i} > ProbCA(e_i = 0)\delta\mu_c \\ \bar{e} & \text{otherwise} \end{cases}$$

I will refer to the following object as the cutoff point, where players are indifferent between the effort choices:

$$c^* \equiv \frac{ProbCA(e_i = 0)\delta\mu_c}{\bar{e}} \tag{1.2}$$

The intuition behind this equation is when choosing whether to exert extra effort or not, there is a tradeoff between incurring the immediate costs of the extra effort for the initial job and reducing future expected costs by reducing her probability of having to undertake corrective action. Below the cutoff point, bidders choose extra effort, whereas above the cutoff point they do not exert

[7]The μ_c in the expected profit equation follows directly from player i's cost draws being independent, coming from the same distribution, and c_{2i} not being realized until corrective action needs to be undertaken. More formally, $E[c_{2i}|c_{1i}] = E[c_{2i}] = \mu_{c_2} = \mu_{c_1} = \mu_c$.

extra effort. When the probability of corrective action increases, the cutoff point also increases, indicating that a higher proportion of bidders will draw costs associated with optimally choosing extra effort. Similarly, when the immediate cost markup of undertaking extra effort $\bar{e}$ increases, c^* decreases, implying that fewer bidders will choose to use extra effort.

Again, letting $\beta(\cdot)$ represent the equilibrium bid strategy, for which the bid amount is invertible and monotonic in the cost draw, I have the following bid function for the case where there is a warranty:[8]

$$b_i = \beta(c_{1i}) = c_{1i} + ProbCA(e_i = 0)\delta\mu_c + \frac{\int_{c_{1i}}^{\bar{c}} [1 - F_{c_1}(x)]^{n-1} dx}{[1 - F_{c_1}(c_{1i})]^{n-1}} \qquad \text{for} \quad c_{1i} > c^* \quad (1.3)$$

$$b_i = \beta(c_{1i}) = (1+\bar{e})c_{1i} + \frac{(1+\bar{e})\int_{c_{1i}}^{c^*} [1 - F_{c_1}(x)]^{n-1} dx}{[1 - F_{c_1}(c_{1i})]^{n-1}} + \frac{\int_{c^*}^{\bar{c}} [1 - F_{c_1}(x)]^{n-1} dx}{[1 - F_{c_1}(c_{1i})]^{n-1}} \qquad \text{for} \quad c_{1i} < c^*$$

1.4 Data and Descriptive Analysis

This section describes the data and provides descriptive evidence for the assumptions underlying the model. Observable data on repair work, which are unique to this data set, are given special attention.

1.4.1 MDOT Highway Procurement Auctions

The data come from the Michigan Department of Transportation (MDOT). Similar to other government entities, MDOT uses sealed-bid, low price procurement auctions in order to allocate projects. In the case of MDOT, they are concerned with the construction and maintenance of state infrastructure, with a focus on roads and bridges.

[8] See Appendix for details.

MDOT makes results from their auctions readily available on their website. The results display the auction ID (which is uniquely assigned to each auction), a brief description of the job, the bidders, their total bids, and also the individual line items for each of the bidders. To elaborate on this last element, when MDOT offers up a project through an auction, they have bidders submit individual dollar estimates for the different components of the project. These items can be removal of past structures, groundwork, installation of barriers, etc. These individual items are then totaled and used to evaluate the competitiveness of the bids.

I first analyze all data for both warranty and standard auctions that fall within the entire window available, 1999-2014. Table 1.1 presents some basic summary statistics related to the auctions and winning bids in both settings. Due to my use of ex post data on corrective action in my empirical setup, I need to limit my sample to make sure the warranty periods have expired by the time I collected the data. Additionally, I limit the start of my data to the time when MDOT began widely adopting the practice of experimenting with warranties, which happened in 2001. Lastly, with a vast majority of my bids coming from auctions with $n = 2, ..., 6$, I will limit my attention to these auctions for reasons that will be more apparent in the estimation section. Table 1.2 presents the summary statistics for the data that I use as my estimation subsample. To get a better sense of the types of work undertaken, Table 1.3 displays the breakdown of jobs by type.

1.4.2 Additional Data on Corrective Action

In addition to having information on the bids and project characteristics, I have information related to the (potential) corrective action undertaken by contractors on all warranty project. These data were obtained from MDOT by the Freedom of Information Act. The data show which projects needed corrective action, when the corrective action started and ended, as well as an estimate by the contractor as to how much the corrective action cost. This estimate of costs is a noisy measure, but the other information is said to be precisely reported.

1.5 Estimation and Results

I develop a multistage estimation procedure in order to recover the latent cost distribution as well as parameters related to effort. I first adopt additional structure on the cost draws that fits the specific context of my empirical investigation. This structure allows me to separate out costs that arise due to common observable factors from those that are due to individual idiosyncracies. Next I recover the full latent distribution of costs directly from observable bids. Lastly, I use the cost distribution to recover the remaining effort-linked parameters that are central to warranty auctions. Once I have fully recovered the latent cost distribution and remaining effort-linked parameters, I conduct policy and counterfactual analysis related to the decision by the state to implement warranty auctions alongside standard auctions.

One hurdle the econometrician comes across in carrying out analysis on empirical auction data is that in practice the items being auctioned are not, in fact, identical. Some projects are more complex, take place at different

times, and have different requirements for completion. Fortunately, all of the heterogeneity in projects just mentioned is captured in the documents provided by MDOT to the bidders (and me).

There are few established methods for dealing with this observed heterogeneity. I follow Balat, Haile, Hong, and Shum (2015) in accounting for these observables by "homogenizing" the bids before carrying out the rest of my empirical strategy. Alternatively, it is possible to extend my nonparametric approach below by conditioning on the covariates when estimating the CDFs and pdfs needed to construct the pseudovalues. One benefit of using such a technique would be that all of the identification results easily extend to this conditional case. Unfortunately, though, the practical implementation of such a technique can be difficult if one has a rich set of covariates, as I do. In my case, the curse of dimensionality causes my pseduovalue calculations to be imprecise.

My approach essentially normalizes the bids by subtracting off the expected mean conditional on covariates. After normalizing the bids through this method, I can directly apply the estimation methods in the case where identical goods are being auctioned. I assume cost draws are made up of two components, as defined by:

$$c_{i\ell} = \Psi(\mathbf{z}_\ell) a_{i\ell} \tag{1.4}$$

where $\Psi(\cdot)$ is some function of observable auction level covariates, $\mathbf{z}_\ell$, and $a_{i\ell}$ is an individual and auction level i.i.d. draw that is independent of $\mathbf{z}_\ell$. In my analysis, I follow Asker (2010) and define $\Psi(\mathbf{z}_\ell) = e^{\psi \mathbf{z}_\ell}$. This allows for both the mean and variance to be affected by covariates, which is realistic in my setting.

Balat, Haile, Hong, and Shum (2015) and Asker (2010) show that if one thinks of there being some normalized value of covariates, $\mathbf{z_0}$, which need not even exist in my data, then:

$$\beta(c_{i\ell}; n, \mathbf{z}_\ell) = \Psi(\mathbf{z}_\ell)\beta(c_{i\ell}; n, \mathbf{z_0})$$

Letting $\alpha(n) = E[ln(\beta(C; n, \mathbf{z_0}))]$,

$$ln(b_{i\ell}) = \alpha(n_\ell) + \psi\mathbf{z}_\ell + \epsilon_{i\ell}$$

Thus by running a standard regression of bids on observable auction covariates, including dummy variables for the number of bidders in each auction, I am able to recover $\hat{\Psi}()$. In order to end up with a "homogonized" bid $b^0_{i\ell}$, I subtract the shift in the bid that comes from the observable covariates:

$$ln(b^0_{i\ell}) = ln(b_{i\ell}) - \hat{\psi}(\mathbf{z}_\ell)$$

The homoginized bid represents the bid the player would have submitted in the case where their observable covariates were $\mathbf{z_0}$. The coefficients from this homogonization regression are presented in Table 1.4. The results appear in line with what I expected having spoken to industry experts.

In what follows, I will use b to represent the homogonized bids and c to represent this i.i.d. cost draws in order to simplify notation and facilitate comparison with other papers in this literature. The seminal work of Guerre, Perrigne, and Vuong (2000) provided nonparametric identification results that apply to my setting, and I follow their constructive estimation procedure using the homogonized bids in order to recover the latent cost distribution.

Guerre, Perrigne, and Vuong (2000) were the first to show that the first order condition in equation 1.1 can be rewritten as:

$$c_{i\ell} = \xi(b_{i\ell}, G, n) \equiv b_{i\ell} - \frac{(1 - G_B(b_{i\ell}; n))}{(n-1)g_B(b_{i\ell}; n)}$$

where I refer to $\xi(b_{i\ell}, G, n)$ as the inverse bid function, since it maps bids and their distribution into cost draws. This inverse bid function is the key development that allows me to recover the latent cost distribution directly from the observed bids. My estimation strategy requires me to first estimate the bid distribution and density using:

$$\hat{G}_B(b; n) = \frac{1}{nT_n} \sum_{t=1}^{T} \sum_{i=1}^{n} 1\{b_{it} \leq b, n_t = n\}$$

and

$$\hat{g}_B(b; n) = \frac{1}{nT_n h_g} \sum_{\ell=1}^{L} \sum_{i=1}^{n_\ell} K\left(\frac{b - b_{i\ell}}{h_g}\right) 1\{n_\ell = n\}$$

Kernel density estimators requires a choice of kernel K and bandwidth h_g. I use the triweight kernel (see Li, Perrigne, and Vuong (2000), Li, Perrigne, and Vuong (2002), and Athey and Haile (2007) for discussions of this selection and its advantageous properties):

$$K(u) = \frac{35}{32}(1 - u^2)^3 I(|u| \leq 1)$$

with the bandwidth selection following the selection criteria of Silverman (1986). That bandwidth is defined as:

$$h_g = 2.978 \times 1.06 \hat{\sigma}_b (nL)^{-1/5}$$

where $\hat{\sigma}_b$ is the standard deviation of observed bids. As has been noted in the literature, nonparametric estimators can be biased near the boundaries, therefore I trim the boundary to eliminate the unwanted properties of the estimator (see Li, Perrigne, and Vuong (2002) for details related to the trimming procedure).

Having recovered both $\hat{G}_B$ and $\hat{g}_B$, I am able to use equation 1.5 to back out an estimate for the cost draw $\hat{c}_{i\ell}$, known as the pseudocost, for each bid $b_{i\ell}$:

$$\hat{c}_{i\ell} = b_{i\ell} - \frac{(1 - \hat{G}_B(b_{i\ell}))}{(n-1)\hat{g}_B(b_{i\ell})}$$

After recovering all of the estimated pseudocosts, I consistently estimate the true underlying cost distribution and density through similar nonparametric estimators. The CDF uses an empirical distribution estimator analogous to the one employed on the bids. For the pdf, I use the following estimator:

$$\hat{f}_c = \frac{1}{nLh_f} \sum_{\ell=1}^{L} \sum_{p=1}^{n} K\left(\frac{c - \hat{c}_{p\ell}}{h_f}\right)$$

with bandwidth

$$h_f = 2.978 \times 1.06 \hat{\sigma}_c (nL_T)^{-1/5}$$

where $\hat{\sigma}_c$ is the standard deviation of pseudocosts after the trimming stage, and L_T represents the number of auctions that remain after trimming.

Figures 1.1a and 1.1b plot $\hat{F}_c$ and $\hat{f}_c$, respectively. I use a block bootstrap to estimate the 95% confidence intervals (Kunsch et al. (1989)). These estimates are enough to explore a wide range of policy relevant questions, but in order to address issues related to effort choice and warranty costs, I need to consider additional data to recover relevant parameters.

I use $\hat{F}_c$, $\hat{f}_c$, and the unique ex post corrective action data to recover the remaining effort-linked parameters. Before proceeding with estimation, I establish the proportion of the enitre project that is typically addressed by contractors when corrective action is needed. Recall that this amount is represented by δ in my model. I carefully consider the ex post self-reported data in coordination with expert insight from MDOT engineers to arrive at an value of $\delta = 8\%$.[9] Next I use the Generalized Method of Moments to match theoretical moments that are functions of effort and the probability of corrective action to their empirical counterparts:

$$Moment\ n : E[ProbCA(e_i = 0) * [1 - F(c^*)]^n - \overline{ProbCA}_n] = 0 \text{ for n = 2,..,N} \tag{1.5}$$

where I define $\overline{ProbCA}_n$ as the fraction of all warranty auctions with n bidders that need corrective action. The Appendix shows how having the ex post corrective action probabilities for two different numbers of bidders and the underlying cost distribution is enough to identify the remaining terms, but I choose to impose the restriction for $n = 2, ..., 5$ to improve the precision of my estimates. In my data, I see whether corrective action needed to be undertaken for each warranty auction. The empirical counterpart to the moment above is:

$$\hat{M}_n = ProbCA(e_i = 0) * [1 - \hat{F}_c(c^*)]^n - \frac{1}{L_n} \sum_{\ell_n=1}^{L_n} 1\{\text{corrective action undertaken}\}_{\ell_n} \tag{1.6}$$

[9] See Appendix for discussion of these data and how I arrived at this value. As a robustness check, I reestimated the model by perturbing δ within reasonable bounds and all of the results remained qualitatively similar.

From these moments, I estimate $\bar{e} = .043$, with standard error of 0.018, and $ProbCA(e_i = 0) = .515$, with standard error of 0.17. Interpreted separately, these estimates imply that immediate costs are increased by a 4.3% when extra effort is used, and when extra effort is not used the probability of corrective action is slightly over 50%. Although these estimates can be discussed separately as I have just done, perhaps the implications from them are best understood when taken together. One direct implication from my estimates is that 29.3% of bidders' cost draws associated with higher effort work.[10] Although this simple statistic may not be a particular target goal set out by policymakers, it can be easily understood by all parties involved with deciding the fate of the warranty auction program. When comparing the immediate costs of the program to the benefits, such a statistic can serve as a helpful reference to recognize how the incentives provided by the warranty contracts induce a substantial portion of the contractors to undertake higher effort work.

As a test of my model's fit, I use the estimated primitives to simulate warranty bids and compare moments to all true warranty bids as well as just the true winning bids. In this simple exercise, I focus on warranty bids where $n = 4$, because this is the average number of bidders in these particular auctions. In addition to considering what my model predicts when calculating moments, I consider what I call the "myopic approach." In the myopic case, I simulate bids from the cost draw using the traditional bid function. This is considered myopic bidding because it represents what the bidder would do if they were only worried about immediate costs and did not take into account future costs. In this case,

[10] Since winning bidders in my setting are the lowest bidders, this percentage is even higher for winners.

they would never exert costly extra effort and bid as if $ProbCA(e_i = 0) = 0$. This myopic case should not be taken too seriously, since we assume bidders do account for the cost tradeoffs of exerting extra effort, but it serves as a nice benchmark for understanding how these new costs affect bids.

Figure 1.2 plots the empirical distribution of observed warranty bids (solid black) versus those generated from my warranty model simulation (dashed red). Table 1.5 compares the first two moments for observed and simulated bids, including the moments from the myopic case. As expected, the myopic case underestimates the average bid. My model's simulated bids slightly exceed observed winning bids, on average, and tend to show more dispersion than observed bids, but overall achieve a reasonable level of fit considering warranty bids were not used in estimation of the primitives. In general, my model strongly outperforms the myopic case when trying to match the true observed bids.

1.6 Policy Implications

The results section discussed some of implications that are easily derived from the estimates, but this section looks to briefly explore more in-depth policy implications as they relate to further actions the government could undertake to encourage or or discourage costly effort. These exercises allow me to quantify the effect warranties and the effort-linked parameters have on auction outcomes, which inform how policymakers should approach their use.

1.6.1 Implementation of All Warranty Auctions

One theme of question that kept coming up during discussions with employees from MDOT was how much it would cost the state to do a full implementation

of warranty auctions. In order to address this, I simulate warranty bids from estimated pseduocosts in standard auctions to compare the winning bids from my simulation to those from the observed data. This addresses how much it *would have* cost the state, but I should note that because bids are a function of observable covariates in addition to cost draws, future projects could cost more or less based on project characteristics.

Taking all standard projects and treating them as warranty projects would have immediately increased the cost to the state by $68.7 million in total over the time frame I consider ($17.2 million annually). This is just the immediate cost increase, though, because the state benefits from having the contractor take care of future work. As a rough estimate for the net effect from the state's perspective, accounting for the fact that warranty auctions relieve the state of having to perform maintenance, I add the expected future costs of corrective action to each of the standard bids (assuming this cost will be directly paid by the state) and compare these "total costs" from standard auctions to the total costs under the warranty auction setting.[11] This approach indicates that the state has a net benefit of $3.88 million annually from using the innovative warranty contracts.[12] It is beyond the scope of this paper to model the state as strategically selecting when to undertake these projects given state funding constraints, but this complication could help explain future contract assignments

[11]The total costs in the warranty auction setting are simply the winning bids, since the state is not responsible for corrective action.

[12]This approach puts some strong assumptions on the state, such that if they were to perform corrective action their cost draws would come from the same distribution as the contractors' cost draws. Since the state is outsourcing the projects to begin with, it can be reasonably argued that the state's cost draws are likely to be higher than the contractors' draws. If that were the case, warranties would prove to be even more effective at saving the state money.

that seem suboptimal given the net benefits discovered here.

1.6.2 Long-Term Reduction in Cost of Effort

Lastly, as was pointed out in the modeling section, states are hoping that warranties encourage innovation and can help lower the costs of increased effort over a longer time horizon. As cost saving technology and techniques are developed and refined as a consequence of warranty auctions, states would reap the benefits of the lower costs. In this section I simulate a set of warranty auctions under different settings of future $\bar{e}$ in order to let policymakers better understand potential longer-term savings of sticking with this program that are not captured over the short span of my data.

Since my data do not provide any insight into the expected cost reductions of new technology, and those I spoke with at MDOT did not feel comfortable going on record with their estimates of future cost reductions, I present a range of results that paint a broad picture of potential future benefits. More specifically, I build on results from the last section where I simulated warranty bids from the set of pseduocosts backed out from the standard auction and repeat the analysis for $\hat{e} = .9\bar{e}, .75\bar{e}$, and $.5\bar{e}$. When thinking about how this hypothetical reduction in $\bar{e}$ affects bids, its important to remember that not only will this reduce immediate costs for those already choosing to undertake higher effort work, it will also encourage more bidders to undertake higher effort work.[13]

From the baseline case in the last section where $\bar{e}$ simply is equal to my original estimate, the immediate cost markups of using warranties over standard

[13]To see how this works mechanically, remember the definition of c^*. In equation 1.2 $\bar{e}$ is in the denominator. Lower $\bar{e}$ leads to higher c^*, which in turn leads to a higher percentage of bidders undertaking extra effort work.

auctions are \$62.4 million, \$55.9 million, and \$37.3 million for the cases of $.9\bar{e}, .75\bar{e}$, and $.5\bar{e}$, respectively. Thus, in the case where the future $\bar{e}$ is reduced by half from its current level, the state can expect to save and additional \$31.2 million of its current immediate costs over the same time frame. While it may be idealistic to assume that there could be a 50% reduction in the effort component over the longer term, it is worth considering since new technologies, like the ones discussed in the introduction, are rapidly being tested and used by more companies.

1.7 Conclusion

This paper considers procurement auctions where the item being procured is under warranty by the winning contracting firm. This contract structure incentivizes firms to potentially use a higher level of effort, as more effort today leads to lower expected future costs. The model I develop explicitly accounts for contractors' latent effort choices affecting both immediate and expected future costs in opposite directions.

I estimate that roughly 30% of bidders are willing to undertake extra effort in order to offset future costs. If all of the standard auctions were to have been conducted as warranty auctions, MDOT would have needed to pay \$68.7 million more in terms of immediate costs, but would have benefited from the risk reduction and lower future costs. I find the state saved as least \$3.88 million annually by implementing these warranty auctions.

One of the major reasons states claim to use warranty auctions is to encourage the development and adoption of better resources and techniques. My

counterfactual evidence shows that if the immediate cost markup associated with undertaking this high effort, $\bar{e}$, falls by 50% due to, say, widespread adoption and learning-by-doing, then the additional immediate cost associated with undertaking these auctions falls to less than $40 million. This decrease is partly due to the reduction in costs associated with those already choosing to use higher effort but also associated with the endogenous shift to using the less expensive higher effort techniques by those currently planning to use regular effort.

Table 1.1: Summary Statistics for All MDOT Auctions 1999-2014

	All Contracts	
	Standard	Warranty
Engineer's Estimate ($M)	1.21	3.35
	(3.3)	(8.9)
Winning Bid ($M)	1.12	3.19
	(3.3)	(8.6)
WinningBid/EngineerEstimate	.930	.956
	(.147)	(.135)
Number of Bidders	5.2	3.69
	(3.1)	(2.0)
Number of Items	63.1	81.7
	(55.7)	(109.3)
N	10906	1775

Note: Means are presented on the first line of each variable row and standard deviations are in parentheses on following line.

Table 1.2: Summary Statistics for Subsampled MDOT Auctions 2001-2004

	All Contracts	
	Standard	Warranty
Engineer's Estimate ($M)	1.05	2.34
	(2.62)	(5.72)
Winning Bid ($M)	.980	2.21
	(2.46)	(5.24)
WinningBid/EngineerEstimate	.934	.968
	(.149)	(.147)
Number of Bidders	3.89	3.30
	(1.32)	(1.26)
Number of Items	55.1	60.4
	(50.8)	(89.3)
N	1,953	477

Note: Means are presented on the first line of each variable row and standard deviations are in parentheses on following line.

Table 1.3: Types of MDOT Projects 2001-2004

	All Contracts	
	Number	Frequency
New Construction and Reconstruction	314	12.92
Rehabilitation and Resurfacing	842	33.65
Preventative Maintenance	398	16.38
Bridge	205	8.43
Traffic, Safety, and Roadside Facilities	336	13.83
Misc.	334	13.78
Total	2,430	100

Table 1.4: Results from Homogenization Regression 2001-2004

Number of Items	0.071
	(0.005)
Year 2002	-0.011
	(0.004)
Year 2003	0.003
	(0.001)
Year 2004	-0.001
	(0.005)
Engineer's Estimate	0.981
	(0.002)
Rehabilitation and Resurfacing	0.003
	(0.001)
Preventative Maintenance	0.003
	(0.005)
Bridge	-0.007
	(0.007)
Traffic, Safety, and Roadside Facilities	0.018
	(0.006)
Misc.	-0.018
	(0.006)

Table 1.5: Moments for Model Fit: n=4

	All Contracts		
	Data	Warr Sim	Myopic Sim
Mean: all bids ($M)	0.92	0.92	0.84
Standard Deviation: all bids ($M)	0.86	0.87	0.80
Mean: winning bids ($M)	0.83	0.87	0.75
Standard Deviation: winning bids ($M)	0.79	0.83	0.72

Table 1.6: Monte Carlo Homogonazation Regression Results

	True Value	Estimate	Standard Error
β_1	0.50	0.50	0.01
β_2	0.30	0.30	0.01

Table 1.7: Monte Carlo Effort Results

	True Value	Mean	Median	Standard Deviation
$\bar{e}$	0.80	0.81	0.80	0.13
$ProbCA$	0.30	0.32	0.31	0.04

Figure 1.1a: CDF of Pseudocosts

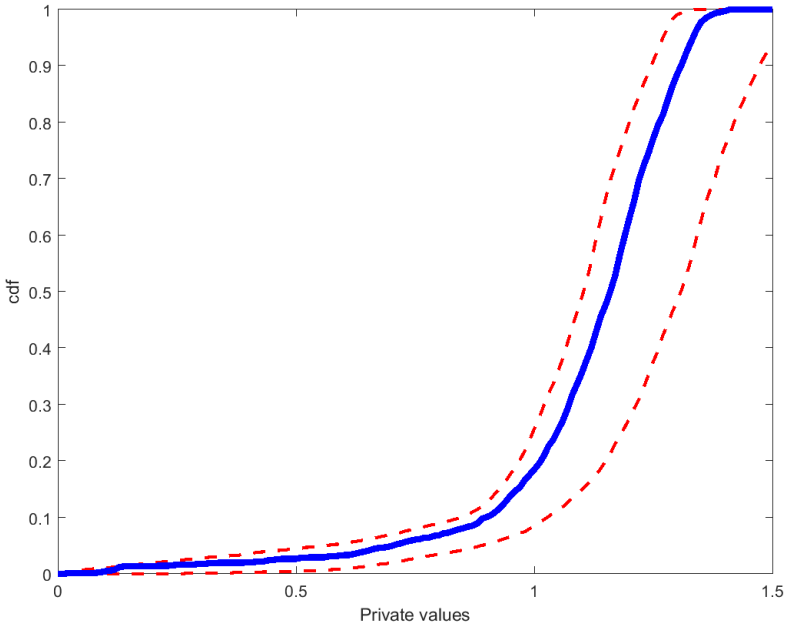


Figure 1.1b: pdf of Pseudocosts

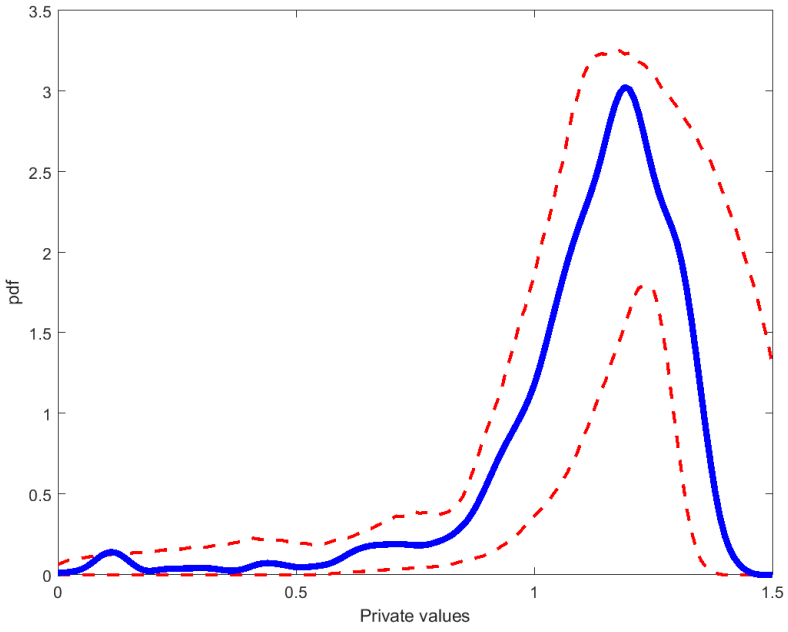
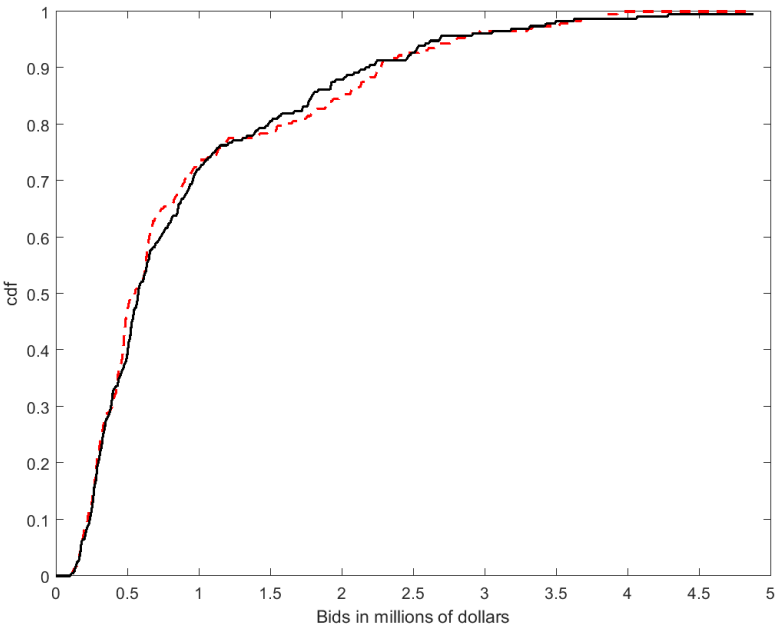
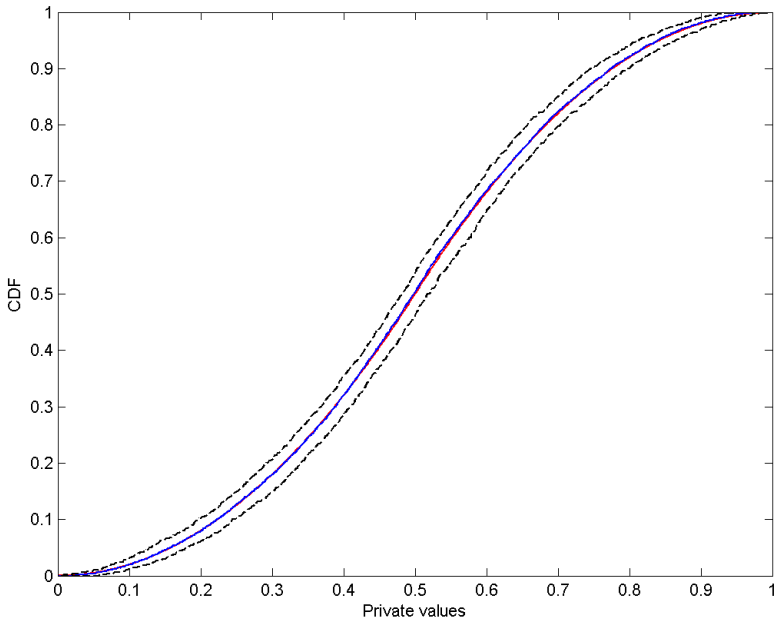


Figure 1.2: Model Fit: CDFs of Observed and Simulated Bids



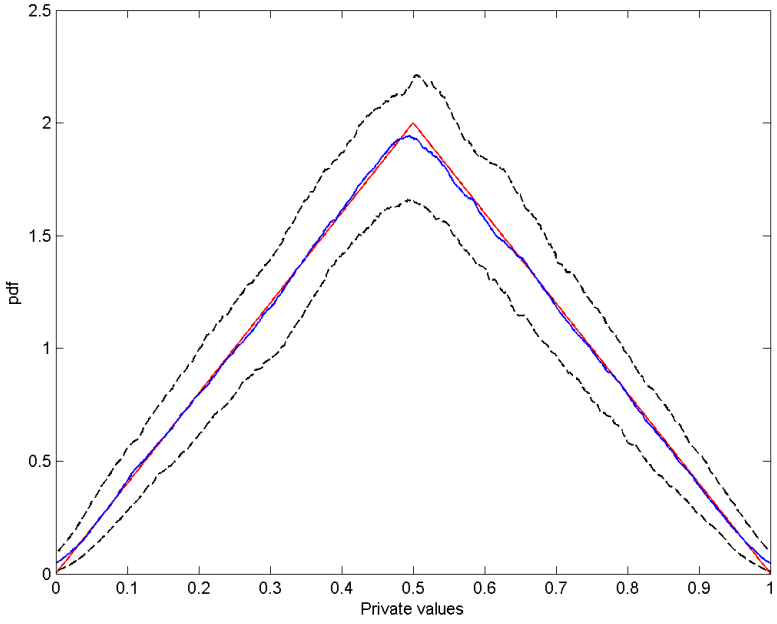
Note: These are the CDFs from the true data (black) and simulated data (red) for warranty auctions with n = 4.

Figure 1.3a: Monte Carlo CDF of True Drawn Costs



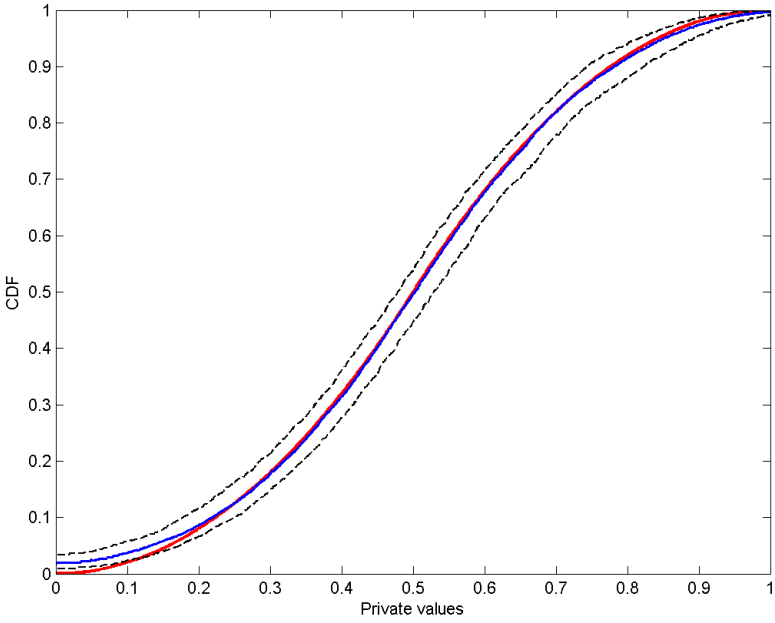
Note: The red line depicts the true (theoretical) CDF, the blue line presents the median of the empirical distribution function estimates performed on the generated cost draws, and the dashed black lines represent the 95% confidence region of the estimates performed on the generated cost draws.

Figure 1.3b: Monte Carlo pdf of True Drawn Costs



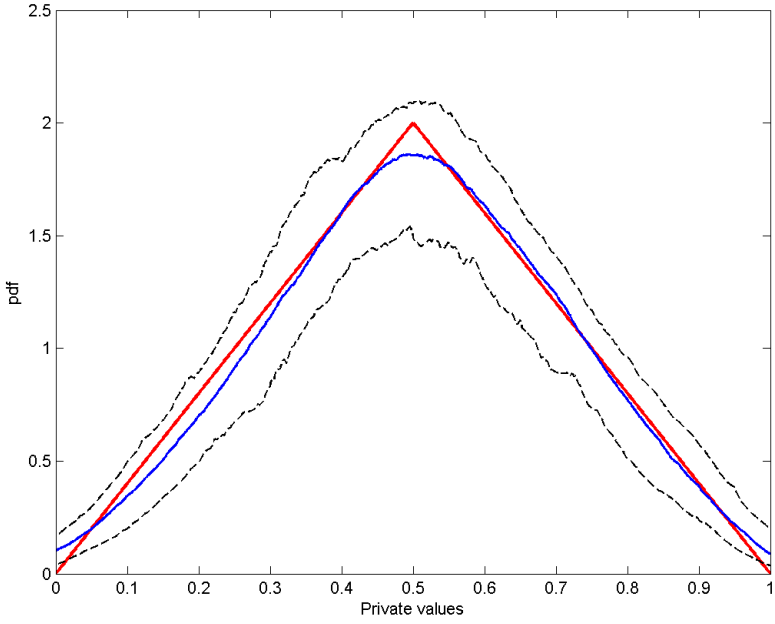
Note: The red line depicts the true (theoretical) pdf, the blue line presents the median of the kernel estimates from the generated cost draws, and the dashed black lines represent the 95% confidence region of the estimates from the generated cost draws.

Figure 1.4a: Monte Carlo CDF of Pseudocosts



Note: The red line depicts the true (theoretical) CDF, the blue line presents the median of the empirical distribution function estimates performed on the estimated pseudocosts, and the dashed black lines represent the 95% confidence region of the estimates performed on the estimated pseudocosts.

Figure 1.4b: Monte Carlo pdf of Pseudocosts



Note: The red line depicts the true (theoretical) pdf, the blue line presents the median of the kernel estimates from the estimated pseudocosts, and the dashed black lines represent the 95% confidence region of the estimates from the estimated pseudocosts.

Chapter 2

On Misclassification Errors in Self-Reported Surveys

2.1 Introduction

Misclassification of drug use is pervasive in self-reported surveys and has important implications for both research and policy. It is well known that estimators ignoring measurement error in the independent variable can be biased and inconsistent, complicating how to interpret results and analysis. Many factors lead to reporting error in population surveys, and the stigma associated with drugs only adds to the difficulty of finding an accurate measure of self-reported drug use. We estimate a nonlinear model with nonclassical measurement error in covariates using data from the National Longitudinal Survey of Youth 1997 in order to find the extent to which measurement error is present in self-reported drug use. Our identification and estimation strategy allow us to consider misclassification in multiple variables, giving us a more realistic framework for analyzing the problem.

Our approach to correct the misclassification error has important implications for many areas of research. There exists a wide body of literature in labor

economics and health economics focusing on the effects of drug use as they relate to youth behavior, wage rates, labor supply, and employment. Kaestner (1994) uses the NLSY to measure the effect of drug use (both marijuana and cocaine) on labor supply. He finds that when looking at a cross section there is a significant negative effect of drug use on labor supply, but when using a longitudinal sample there is no significant effect. MacDonald and Pudney (2000) use two cross sections from the British Crime Survey data to study the relationship between drug use and labor market outcomes. Other studies by DeSimone (2002), French, Roebuck, and Alexandre (2001), and French, Zarkin, Mroz, and Bray (1998) all rely on self-reported data when estimating their results related to drug use and employment. Some of the labor economics literature addresses issues of endogeneity of drug use, but the literature has not given the same attention to measurement error that could be significantly affecting the results.

Some studies from the health economics literature are concerned with how drug use is related to educational attainment and abuse of other substances. Bray, Zarkin, Ringwalt, and Qi (2000) use longitudinal data on students in the US public schools to estimate the relationship between marijuana use and dropping out of school. They find that students using marijuana are roughly 2.3 times more likely to drop out than students who abstain. The authors also investigate how these odds of dropping out vary across the age of the students. A study by Beenstock and Rahav (2002) uses a self-reported measure of drug use to test the Gateway Theory of drug use, and it concludes that the use of cigarettes leads to marijuana use. Norton, Lindrooth, and Ennett (1998) attempt to account for endogenously selected peer groups when analyzing peer effects on substance abuse. Using longitudinal data from a drug use prevention

program, they find a significant role for peer effects even after controlling for peer selection. It would be interesting to investigate how the existing results change after correcting the misreporting error in self-reported drug use.

Our identification strategy is also important to the econometrics literature. Nonlinear modeling with measurement error has been an active area of research recently (see Chen, Hong, and Nekipelov (2011) for a survey). The literature focusing on classical error, where the measurement error is independent of the latent variable, has built off the work by Hausman, Newey, Ichimura, and Powell (1991). Recent studies include Schennach (2004) and Schennach (2007), which use an instrumental variable (IV) approach also featured in this paper. The work by Hui and Walter (1980), Mahajan (2006), Hu (2008), and Hu and Schennach (2008) all relax the assumption of independence between the latent true variable and measurement error. The results of Hu (2008) have recently been used by Balat (2012), Sasaki (2011), and An, Hu, and Shum (2010). Moreover, Molinari (2010) discusses the identification of treatment effect when some observations on treatment are missing. A more closely related work is by Kreider, Pepper, et al. (2011) in which the misreporting of drug use is analyzed under the mixing outcome sampling assumption. This paper uses the results of Hu (2008) in identifying and estimating our model with misclassification, and is the first to estimate such a complicated misclassification error. We allow for two latent variables when estimating the joint distribution of reported and actual drug use, and our method is well suited to be adopted by researchers who find it realistic to have both present and lagged latent variables affect misclassification.

Misreporting errors in self-reported drug use surveys have been extensively studied. Mensch and Kandel (1988) compare the self-reported drugs use in

the 1984 NLSY survey to other surveys and conclude that the NLSY data was subject to underreporting. Some studies attempt to quantify the measurement error in drug use responses using advanced statistical techniques, but none use the methodology presented in this paper. Biemer and Wiesen (2002) employ latent class analysis in order to characterize the classification error in the US National Household Survey on Drug Abuse. Biemer and Witt (1996) use techniques that require repeated measures of the variable of interest, largely focusing on methods first developed by Hui and Walter (1980). Taking as given the findings from medical journals on the ties between marijuana use and increases in health problems (see Polen, Sidney, Tekawa, Sadler, and Friedman (1993) and Volkow, Baler, Compton, and Weiss (2014)), our paper adds to the misreporting literature by offering a way to measure this misclassification while flexibly allowing for multiple variables to have measurement error.

In our model, we assume a nonparametric distribution of misclassification errors which are allowed to be correlated with explanatory variables. With relatively few restrictions we are able to identify the misclassification probabilities and directly estimate latent variables whenever instrumental variables are available by using an eigen-decomposition method. We show that in general, when people indeed use drugs during the current or previous periods, the probability for him or her to misreport is significantly larger than zero. Also, this misreporting probability is contingent on the past misreporting behaviors as well as drug use history.

The rest of the paper is organized as follows. Section 2 describes our model and identification strategy that allows for two latent variable. Section 3 discusses simulation results employing our identification strategy. Section 4 describes data

we use in our estimation and some assumptions we make regarding key variables. This section also presents our estimation results. Section 5 concludes.

2.2 Model and Identification

In this section, we use the nonparametric identification method from Hu (2008) to estimate the misclassification error of reported marijuana use. Specifically, we are interested in the conditional distribution of reported marijuana use, D_t, on the unobserved latent variables indicating the true marijuana use status, D^*, namely $\Pr(D_t|D^*)$.[1] We restrict our attention to discrete measures of all variables. We present results using two approaches: the basic approach using data from four consecutive periods and assuming misclassification of only one variable, and a more general approach using data from five consecutive periods that allows for misclassification of multiple variables. Specific assumptions for each case will be illustrated and discussed in this section. Lastly, two major theorems will be presented which directly address the identification and estimation of our model.

2.2.1 Basic Approach

In this part, we consider five types of discrete variables: the reported health condition variable at time t, H_t, the self-reported drug use at time t, D_t, the latent true level of drug use at time t, D_t^*, the self-reported drug use at time $t-2$, D_{t-2}, and lastly the two other independent variables, (D_{t-1}, H_{t-1}). They are displayed in Table 2.1.

[1] D^* can be a scalar or a vector. In our basic model, D^* simply refers to the true drug use status D_t^*. In our general case, it is a vector indicating the true drug use status both at time t and $t-1$.

We are interested in the misreporting error distribution $\Pr(D_t|D_t^*, W_t)$ where W_t refers to all the other covariates. In order to derive the major theorem for identification and estimation, let us make the following assumptions:

Assumption 1.1

$$\Pr\left(H_t|D_t^*, D_t, D_{t-1}, H_{t-1}, D_{t-2}\right) = \Pr\left(H_t|D_t^*, D_{t-1}, H_{t-1}\right) \tag{2.1}$$

Assumption 1.1 states that conditional on the last-period report about drug use and health condition, D_t and D_{t-2} provide no relevant information beyond D_t^* to predict the current period health condition. We are implicitly assuming that after considering the report from the previous period, the misclassification error is completely independent of H_t. This immediately indicates that the bias of misreporting current-period marijuana use is the same for healthy and unhealthy people alike if they report the same health and drug use status in the previous period.

Next we impose the conditional independence restrictions on the misclassification error.

Assumption 1.2

$$\Pr\left(D_t|D_t^*, D_{t-1}, H_{t-1}, D_{t-2}\right) = \Pr\left(D_t|D_t^*, D_{t-1}\right) \tag{2.2}$$

This assumption states the misclassification error is independent of D_{t-2} and H_{t-1}, conditional on the true latent drug use status, D_t^*, as well as last-period's

reported drug use status, D_{t-1}. This assumption is reasonable because misreporting of marijuana use should only be dependent on whether, in fact, they used drugs this period and how they reported drug use last period as captured by D_{t-1}.

Note that we only assume conditional independence between the misclassification error and (H_t,D_{t-2},H_{t-1}); other than this, we do not restrict the independence between the error and any other covariates such as gender, education level, and marital status. Furthermore, these assumptions are weak in the sense that we do not impose any specific functional forms on the error term.

As mentioned earlier, we focus on discrete cases where for $\tau \in \{t-2, t-1, t\}$:

$$D_\tau = \begin{cases} 1 & \text{if report using marijuana at least once in the period or "no response"} \\ 0 & \text{if report no marijuana use during the period} \end{cases}$$

and the latent variables are:

$$D^*_\tau = \begin{cases} 1 & \text{if used marijuana at least once during the period} \\ 0 & \text{otherwise} \end{cases}$$

And for the dependent variable:

$$H_\tau = \begin{cases} 1 & \text{if in good health} \\ 0 & \text{if in bad health} \end{cases}$$

for any time period τ. Under Assumptions 1.1 and 1.2, one can show that for any function $\omega(\cdot)$

$$\sum_{H_t} \omega(H_t) \Pr(D_t, H_t, D_{t-1}, H_{t-1}, D_{t-2})$$

$$= \sum_{D^*_t} \Pr(D_t | D^*_t, D_{t-1}) \times E[\omega(H_t) | D^*_t, D_{t-1}, H_{t-1}] \times \Pr(D^*_t, D_{t-1}, H_{t-1}, D_{t-2}) \tag{2.3}$$

These equations relate observed distributions to the underlying latent distributions, and they will be used to prove identification of our model.

2.2.1.1 Identification

We define for any given d_{t-1}, h_{t-1},

$$L_{\omega(H_t)(D_t,H_t,d_{t-1},h_{t-1},D_{t-2})} =$$

$$\left[\begin{array}{cc} \sum_{H_t} \omega(H_t) \Pr(D_t = 1, H_t, d_{t-1}, h_{t-1}, D_{t-2} = 1) & \sum_{H_t} \omega(H_t) \Pr(D_t = 1, H_t, d_{t-1}, h_{t-1}, D_{t-2} = 0) \\ \sum_{H_t} \omega(H_t) \Pr(D_t = 0, H_t, d_{t-1}, h_{t-1}, D_{t-2} = 1) & \sum_{H_t} \omega(H_t) \Pr(D_t = 0, H_t, d_{t-1}, h_{t-1}, D_{t-2} = 0) \end{array} \right]$$

$$L_{D_t|D_t^*,d_{t-1}} = \left[\Pr(D_t = i|D_t^* = j, d_{t-1})\right]_{i,j\in\{1,0\}},$$

$$D_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}} = \left[\begin{array}{cc} E[\omega(H_t)|D_t^* = 1, d_{t-1}, h_{t-1}] & 0 \\ 0 & E[\omega(H_t)|D_t^* = 0, d_{t-1}, h_{t-1}] \end{array} \right],$$

$$L_{D_t^*,d_{t-1},h_{t-1},D_{t-2}} = \left[\Pr(D_t^* = i, d_{t-1}, h_{t-1}, D_{t-2} = j)\right]_{i,j\in\{1,0\}},$$

and

$$L_{D_t,d_{t-1},h_{t-1},D_{t-2}} = \left[\Pr(D_t = i, d_{t-1}, h_{t-1}, D_{t-2} = j\right]_{i,j\in\{1,0\}}.$$

We may then show that equation (2.3) is equivalent to

$$L_{\omega(H_t)(D_t,H_t,d_{t-1},h_{t-1},D_{t-2})} = L_{D_t|D_t^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}} \times L_{D_t^*,d_{t-1},h_{t-1},D_{t-2}} \tag{2.4}$$

and corresponding to a degenerated $\omega(\cdot) = 1$

$$L_{D_t,d_{t-1},h_{t-1},D_{t-2}} = L_{D_t|D_t^*,d_{t-1}} \times L_{D_t^*,d_{t-1},h_{t-1},D_{t-2}}. \tag{2.5}$$

We need to make an assumption on an observable matrix in order to allow us to proceed with our eigen-decomposition technique.

Assumption 1.3 The matrix $L_{D_t,d_{t-1},h_{t-1},D_{t-2}}$ is invertible.

Given our assumptions, we invert both sides of (2.5) and multiply these by the corresponding sides of (2.4) to get:

$$\begin{aligned} & L_{\omega(H_t)(D_t,H_t,d_{t-1},h_{t-1},D_{t-2})} \times L^{-1}_{D_t,d_{t-1},h_{t-1},D_{t-2}} \\ & = L_{D_t|D_t^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}} \times L_{D_t^*,d_{t-1},h_{t-1},D_{t-2}} \times \\ & L^{-1}_{D_t^*,d_{t-1},h_{t-1},D_{t-2}} \times L^{-1}_{D_t|D_t^*,d_{t-1}} \\ & = L_{D_t|D_t^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}} \times L^{-1}_{D_t|D_t^*,d_{t-1}} \end{aligned} \tag{2.6}$$

for any given (d_{t-1}, h_{t-1}). For the remainder of this subsection we will refer to $L_{\omega(H_t)(D_t,H_t,d_{t-1},h_{t-1},D_{t-2})} \times L^{-1}_{D_t,d_{t-1},h_{t-1},D_{t-2}}$ as the left-hand side (LHS) matrix. Looking at equation 2.6, we see that the right-hand side is in the form of an eigen-decomposition of the LHS matrix. Thus, each column of the $L_{D_t|D_t^*,D_{t-1}}$ matrix is a eigenvector of the LHS matrix. The diagonal elements of the $D_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}}$ are the corresponding eigenvalues. Therefore we can directly identify the distribution of misclassification errors, $L_{D_t|D_t^*,D_{t-1}}$, from this matrix diagonalization process.

We can see that all the LHS variables are observable from the data, which means the LHS matrix can be directly calculated. In order to complete our identification, we need to place an ordering on the eigenvectors and guarantee uniqueness of the eigenvalues.

Assumption 1.4 For any given d_{t-1},

$$\Pr\left(D_t = 1 | D_t^* = 0, D_{t-1} = d_{t-1}\right) < \Pr\left(D_t = 1 | D_t^* = 1, D_{t-1} = d_{t-1}\right) \tag{2.7}$$

This means the element of the upper-right corner of the misclassification matrix $L_{D_t|D_t^*,d_{t-1}}$ should be smaller than that of the upper-left corner. This assumption is reasonable because the probability of those who do not use drugs but report using drugs during the period should be very small. This assumption has the flavor or truth-telling that will also be present in our general model. Hence, by checking this criteria we can determine the correct order of the eigenvectors along with corresponding eigenvalues.

Lastly, we need to impose one additional assumption on the eigenvalues:

Assumption 1.5 For any given d_{t-1} there exists an h_{t-1} and a function $\omega(\cdot)$ such that,

$$E[\omega(H_t)|D_t^* = i, d_{t-1}, h_{t-1}] \neq E[\omega(H_t)|D_t^* = j, d_{t-1}, h_{t-1}], \text{for any } i \neq j. \tag{2.8}$$

This assumption ensures that the two eigenvalues are not identical to each other. Without this assumption we cannot successfully identify the misclassification matrix because it could be singular, and there would be no variation in the conditional distribution of $D_t|D_t^*$. Note that here we do not require this inequality to hold for each pair of (d_{t-1}, h_{t-1}). Instead, we only need one subgroup of people sharing the same h_{t-1} for any given d_{t-1}, such that their health-related function ω is contingent on true latent drug use at t.

So far we have made a series of assumptions, and the following theorem justifies our identification and estimation:

Theorem 1.1 *Suppose Assumptions 1.1-1.5 hold, then the misclassification probability* $\Pr(D_t|D_t^*, D_{t-1})$ *is nonparametrically identifiable and directly*

estimable.

Remark: A nice feature of our approach is that, the identification result is independent of the potential misclassification of the dependent variable, H_t, so long as its conditional distribution satisfies Assumption 1.4 and 1.5.

This basic approach uses an eigen-decomposition technique and identifies 2×2 misclassification matrices. The next section will cover a more complicated model, where we identify and estimate 4×4 misclassification matrices.

2.2.2 General Approach

Now we use a more general approach, where we impose assumptions that are empirically more reasonable than those in the first model. Namely, we generalize our conditional independence assumptions by also incorporating a latent variable from a previous period. We still have in total five types of discrete variables, yet in each type, we include more variables. The detailed description is illustrated in Table 2.2.

Here, we use conditional probability $\Pr(D_{t+1}, D_t | D_t^*, D_{t-1}^*, W_t)$ to describe the misreporting behavior, where W_t refers to all the other covariates. Similar to the basic case, we make the following assumptions for identification and estimation:

Assumption 2.1

$$\Pr\left(H_t | D_t^*, D_{t-1}^*, D_{t+1}, D_t, D_{t-1}, D_{t-2}, D_{t-3}, H_{t-1}\right) = \Pr\left(H_t | D_t^*, D_{t-1}^*, D_{t-1}, H_{t-1}\right) \tag{2.9}$$

This assumption indicates that the proxy drug use variables, (D_{t+1}, D_t), and the instrumental variables, (D_{t-2}, D_{t-3}), do not tell us anything more useful about the person's current health condition as long as we know the actual two-period drug use status and their previous-period reported health and drug use. In other words, the misclassification error, conditional on $\left(D_t^*, D_{t-1}^*\right)$ and (D_{t-1}, H_{t-1}), is independent of H_t. Compared with our previous assumption in the basic approach, now we allow the current-period health condition to rely on both current-period true drug use and true drug use from the previous period. This assumption is more realistic in the sense that people's health condition is usually related to his or her drug use history.

We now impose restrictions on the conditional independence between the misclassification error and the instrument variable.

Assumption 2.2

$$\Pr\left(D_{t+1}, D_t | D_t^*, D_{t-1}^*, D_{t-1}, D_{t-2}, D_{t-3}, H_{t-1}\right) = \Pr\left(D_{t+1}, D_t | D_t^*, D_{t-1}^*, D_{t-1}\right) \tag{2.10}$$

Assumption 2.2 implies that the misclassification error is contingent on actual drug use over two periods. We assume people's misreporting decisions only depend on whether they used drugs and how they reported previously. This assumption is more general when compared to that in the basic model, and this can be illustrated in a simple example. Consider a person who used marijuana last year but did not report using it, which is represented by $D_{t-1}^* = 1, D_{t-1} = 0$.

Given that he or she is using marijuana this year, $D_t^* = 1$, his or her probability of misreporting could be higher than those who did not use drug and did not report last year, but indeed use drug this year, $D_{t-1}^* = 0, D_{t-1} = 0, D_t^* = 1$. This can be justified by assuming the former person has a "habit" for misreporting his or her drug use, whereas the latter one is more likely to tell the truth for both periods. On the other hand, there is another driving force in the opposite direction. The former person who has been using drugs for two years might be more likely to report truthfully since he identifies himself as a "frequent" drug user; the latter person who just switched to marijuana this year might be less likely to report this drug use out of fear of getting into trouble. These two forces makes the conditional probabilities $\Pr\left(D_t | D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 0, h_{t-1}\right)$ and $\Pr\left(D_t | D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 0, h_{t-1}\right)$ hard to compare, but we have to separate them. In our previous model, however, we assume these two probabilities are the same regardless of an individual's behavior the previous year.

Furthermore, Assumption 2.2 is weaker than Assumption 1.2 since in this model we assume the misclassification error is independent of (D_{t-2}, D_{t-3}), given true latent variables $\left(D_t^*, D_{t-1}^*\right)$ and two extra observable variables. This is implied by Assumption 1.2 where we assume the error is independent of $D_{\tau-2}$ given D_τ^* and other observable variables, for all τ.

Under Assumptions 2.1 and 2.2, we may have

$$\sum_{H_t} \omega\left(H_t\right) \Pr\left(D_{t+1}, D_t, H_t, D_{t-1}, H_{t-1}, D_{t-2}, D_{t-3}\right)$$

$$= \sum_{D_t^*, D_{t-1}^*} \Pr\left(D_{t+1}, D_t | D_t^*, D_{t-1}^*, D_{t-1}\right) \times E\left[\omega\left(H_t\right) | D_t^*, D_{t-1}^*, D_{t-1}, H_{t-1}\right] \times$$

$$\Pr\left(D_t^*, D_{t-1}^*, D_{t-1}, H_{t-1}, D_{t-2}, D_{t-3}\right). \tag{2.11}$$

Equation (2.11) relates observed distributions to the underlying latent distributions, and they will be used to prove identification of our model.

2.2.2.1 Identification

Again, we define matrices which are analogous to the components of equations

$$L_{\omega(H_t)(D_{t+1}, D_t, H_t, d_{t-1}, h_{t-1}, D_{t-2}, D_{t-3})}$$

$$= \begin{bmatrix} g_1(1,1,1,1) & g_1(1,1,1,0) & g_1(1,1,0,1) & g_1(1,1,0,0) \\ g_1(1,0,1,1) & g_1(1,0,1,0) & g_1(1,0,0,1) & g_1(1,0,0,0) \\ g_1(0,1,1,1) & g_1(0,1,1,0) & g_1(0,1,0,1) & g_1(0,1,0,0) \\ g_1(0,0,1,1) & g_1(0,0,1,0) & g_1(0,0,0,1) & g_1(0,0,0,0) \end{bmatrix}$$

where $g_1(i,j,r,k) = \sum_{H_t} \omega\left(H_t\right) \Pr\left(D_{t+1} = i, D_t = j, H_t, D_{t-1}, H_{t-1}, D_{t-2} = r, D_{t-3} = k\right)$, for any i, j, r, k;

$$L_{D_{t+1}, D_t, d_{t-1}, h_{t-1}, D_{t-2}, D_{t-3}}$$

$$= \begin{bmatrix} g_2(1,1,1,1) & g_2(1,1,1,0) & g_2(1,1,0,1) & g_2(1,1,0,0) \\ g_2(1,0,1,1) & g_2(1,0,1,0) & g_2(1,0,0,1) & g_2(1,0,0,0) \\ g_2(0,1,1,1) & g_2(0,1,1,0) & g_2(0,1,0,1) & g_2(0,1,0,0) \\ g_2(0,0,1,1) & g_2(0,0,1,0) & g_2(0,0,0,1) & g_2(0,0,0,0) \end{bmatrix}$$

where $g_2(i,j,r,k) = \Pr(D_{t+1} = i, D_t = j, d_{t-1}, h_{t-1}, D_{t-2} = r, D_{t-3} = k)$,for any i, j, r, k;

$$L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}}$$

$$= \begin{bmatrix} g_3(1,1|1,1) & g_3(1,1|1,0) & g_3(1,1|0,1) & g_3(1,1|0,0) \\ g_3(1,0|1,1) & g_3(1,0|1,0) & g_3(1,0|0,1) & g_3(1,0|0,0) \\ g_3(0,1|1,1) & g_3(0,1|1,0) & g_3(0,1|0,1) & g_3(0,1|0,0) \\ g_3(0,0|1,1) & g_3(0,0|1,0) & g_3(0,0|0,1) & g_3(0,0|0,0) \end{bmatrix}$$

where $g_3(i,j|r,k) = \Pr(D_{t+1} = i, D_t = j|D_t^* = r, D_{t-1}^* = k, d_{t-1})$, for any i, j, r, k;

$$D_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}}$$

$$= \begin{bmatrix} E[1,1] & & & \\ & E[1,0] & & \\ & & E[0,1] & \\ & & & E[0,0] \end{bmatrix}$$

where, $E[i,j] = E_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}}[i,j] = E\left[\omega(H_t)|D_t^* = i, D_{t-1}^* = j, d_{t-1}, h_{t-1}\right]$, for any i, j;

$$L_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}}$$

$$= \begin{bmatrix} g_4(1,1,1,1) & g_4(1,1,1,0) & g_4(1,1,0,1) & g_4(1,1,0,0) \\ g_4(1,0,1,1) & g_4(1,0,1,0) & g_4(1,0,0,1) & g_4(1,0,0,0) \\ g_4(0,1,1,1) & g_4(0,1,1,0) & g_4(0,1,0,1) & g_4(0,1,0,0) \\ g_4(0,0,1,1) & g_4(0,0,1,0) & g_4(0,0,0,1) & g_4(0,0,0,0) \end{bmatrix}$$

where $g_4(i,j,r,k) = \Pr(D_t^* = i, D_{t-1}^* = j, d_{t-1}, h_{t-1}, D_{t-2} = r, D_{t-3} = k)$, for any i, j, r, k. Thus, the matrix notation would be written as:

$$L_{\omega(H_t)(D_{t+1},D_t,H_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3})} = L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}} \times L_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} \quad (2.12)$$

and corresponding to a degenerated $\omega(\cdot) = 1$

$$L_{D_{t+1},D_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} = L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}} \times L_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} \tag{2.13}$$

We need to make an assumption on an observable matrix in order to allow us to proceed with an eigen-decomposition technique.

Assumption 2.3 The matrix $L_{D_{t+1},D_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3}}$ is invertible.

Given these assumptions, we invert both sides of (2.13) and multiply these by the corresponding sides of (2.12) to obtain:

$$\begin{aligned}
&L_{\omega(H_t)(D_{t+1},D_t,H_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3})} \times L^{-1}_{D_{t+1},D_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} \\
=&L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}} \times L_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} \times \\
&L^{-1}_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} \times L^{-1}_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}} \\
=&L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}} \times L^{-1}_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}}
\end{aligned} \tag{2.14}$$

As was done with the basic model, we will refer to $L_{\omega(H_t)(D_{t+1},D_t,H_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3})} \times L^{-1}_{D_{t+1},D_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3}}$ as the LHS matrix. Thus, the $L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}}$ matrix is the eigenvector matrix of the LHS matrix, and the diagonal elements of the $D_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}}$ are the corresponding eigenvalues. We can directly identify the distribution of misclassification errors, or $L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}}$, from this matrix diagonalization process.

As discussed in previous section, after recovering the eigenvectors and eigenvalues, we need to determine the correct ordering that is consistent with the LHS matrix. Therefore we make the following assumption:

Assumption 2.4 For $D_{t-1} = 1$,

$$Pr(D_t = 1|D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 1) > Pr(D_t = 1|D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 1)$$
$$> Pr(D_t = 1|d_t^*, d_{t-1}^*, D_{t-1} = 1) \text{ for other } (d_t^*, d_{t-1}^*),$$

and

$$E\left[H_t|D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 1\right] > E\left[H_t|D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 1\right].$$

For $D_{t-1} = 0$,

$$Pr(D_t = 0|D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0) > Pr(D_t = 0|D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 0)$$
$$> Pr(D_t = 0|d_t^*, d_{t-1}^*, D_{t-1} = 0) \text{ for other } (d_t^*, d_{t-1}^*),$$

and

$$E\left[H_t|D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 0\right] > E\left[H_t|D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 0\right].$$

This assumption directly gives us the ordering of the eigenvalues and eigenvectors needed to establish identification. For $D_{t-1} = 1$ case, we assume people primarily consider their current-period true drug use status as a reference of how they report it. Specifically, we assume that people who are using marijuana in current period would be more likely to report using it than would people who are not currently using marijuana. Another factor that influences people's current reporting behavior is their last-period misreporting behavior. We assume people who have been using marijuana for both periods and reported using it last period have the highest probability of reporting this period again; and people who use marijuana this period but not last period, but report using it last period, have the second highest probability of reporting drug use this period. Thus, we are able to identify two columns of our misclassification matrix, and for the other two we use eigenvalues to distinguish them, as is illustrated in

the second equation for $D_{t-1} = 1$ case. We assume that the expected reported health condition is better for those who do not use marijuana either period, compared with those who use marijuana last period but not this period. Similarly, for $D_{t-1} = 0$ case, we assume that people who do not use marijuana in either period have the highest probability of not reporting marijuana use this period. People who do not use marijuana this period but used it last period would have the second highest probability of not reporting it now. For the other two cases, we assume the expected health condition for those who use marijuana both periods is worse than those who only used it in the current period.[2] These assumptions have the flavor of "truth telling" as a eigenvalue and eigenvector ordering mechanism.

Lastly, to make the identification effective, we need to impose restrictions on the eigenvalues:

Assumption 2.5 For all values of D_{t-1} there exists some h_{t-1} such that

$$E\left[\omega\left(H_t\right)|D_t^*, D_{t-1}^*, d_{t-1}, h_{t-1}\right] \neq E\left[\omega\left(H_t\right)|D_t^*, d_{t-1}, h_{t-1}\right] \neq E\left[\omega\left(H_t\right)|D_{t-1}^*, d_{t-1}, h_{t-1}\right]. \tag{2.15}$$

It states that for those who reported using drugs last period, there exists at least one subgroup of people who share (d_{t-1}, h_{t-1}) for whom their current health condition depends both on current drug use status and previous drug use. Otherwise, we cannot effectively tell apart the difference of the misclassification errors between $\left(D_t^* = i, D_{t-1}^* = j\right)$ type people and $\left(D_t^* = i, D_{t-1}^* = k\right)$

[2] See Polen, Sidney, Tekawa, Sadler, and Friedman (1993) and Volkow, Baler, Compton, and Weiss (2014) for further motivation of these assumptions.

type people when $j \neq k$. In that case we would have duplicates in the eigenvalues, and the identification fails. The following theorem justifies our identification and estimation:

Theorem 2.1 *Suppose Assumptions 2.1-2.5 hold, then the conditional probability* $\Pr\left(D_{t+1}, D_t | D_t^*, D_{t-1}^*, D_{t-1}\right)$ *is nonparametrically identifiable and directly estimable.*

Remark: As is discussed in the basic approach, the misclassification of the dependent variable H_t has no effect on our identification results, so long as the conditional distribution satisfies Assumption 2.4 and 2.5.

In the next section we run simulation for both the 2×2 and the 4×4 cases to justify the validity of our models, and then we use NLSY data to estimate the misclassification errors in marijuana use.

2.3 Simulation

2.3.1 Basic Approach

In this section we generate a set of data using underlying parametric values, and then we use our nonparametric identification method to estimate the misclassification matrices. Information regarding the data generation process can be found in the appendix. We compare these estimates with the true underlying matrices in order to validate our method. We first generate all the data that are related to our estimation, namely $(D_t^*, H_t, H_{t-1}, D_t, D_{t-1}, D_{t-2})$. Using both observable data, $(H_t, H_{t-1}, D_t, D_{t-1}, D_{t-2})$, and unobservable latent variable D_t^*, we are able to calculate the sample average of the misclassification matrices of

interest, $\Pr(D_t|D_t^*, D_{t-1}, H_{t-1})$. Lastly, we use only observable data to estimate $\widehat{\Pr}(D_t|D_t^*, D_{t-1}, H_{t-1})$ according to Theorem 1.1. If the difference of these two matrices converges as the sample size increases, then our basic model is valid.

2.3.1.1 Simulation Results

The mean, median, and standard errors for simulation results are displayed in Tables 2.3. There are several things to point out here. Firstly, the means and medians are converging to the true value as the sample size increases. Secondly, the standard errors are decreasing as the sample size increases. This convergence indicates that our method of identification is at least correct asymptotically. In fact when the sample size is 7000, which is very close to that of our real data, the simulation results are already significant. This further validates our estimation using the basic model.

The results indicate that some estimated probabilities are much more accurate than others. This is due to the loss in accuracy associated with inverting the LHS matrix if it is near singular. This loss in accuracy can be remedied, however, by increasing sample size and therefore canceling out inaccurate estimates from a single trial.

2.3.2 General Approach

In the general model, we need more data regarding previous drug use and health conditions. We require data on $(H_t, H_{t-1}, D_t^*, D_{t-1}^*, D_{t+1}, D_t, D_{t-1}, D_{t-2}, D_{t-3})$. We use the data generation process in the next subsection to generate all these data, then we use all information, observable and unobservable, to calculate the sample mean of the true values of the misclassification matrices. Lastly,

as before, we only use observable variables to estimate this matrix according to Theorem 2.1 and compare this estimation result with the underlying true values.

This simulation process, compared with that in the basic model, is more complex and needs careful consideration. Instead of estimating two 2×2 matrices, we are now estimating two 4×4 matrices. As was the case in the basic model, any inaccuracy can be resolved as we increase the sample size. These features will be discussed in the simulation results section.

2.3.2.1 Simulation Results

The mean, median, and standard errors for simulation results are displayed in Tables 2.4. We can see that when we have a sample size of 7000 and 10000, the estimation results are farther from the true values when compared to the basic model. The standard errors are larger than those found in the basic model as well. Our results once again show that our general model is also correct at least asymptotically, or when we have relatively large sample size.

2.4 Estimation

2.4.1 Data

The data we use in our analysis come from the National Longitudinal Survey of Youth 1997 (NLSY97). The NLSY97 is a panel survey conducted by the Bureau of Labor Statistics that captures the transition from youth to adulthood. The data consist of detailed information on a cohort of approximately

9000 youths who were between the ages of 12 and 16 when the survey was first administered in 1997. These individuals were asked questions covering areas such as background, employment, drug and alcohol use, health, family, and education. Previous literature that has investigated self-reported drug use sometimes focused on the National Household Survey on Drug Abuse data, but we take advantage of the panel nature of the NLSY97 to obtain our estimates of misclassification.

The variables central to our study include marijuana use and health condition. Additionally, we can account for gender, age, marital status, ethnicity, and education level, but these are not crucial to our estimation and identification of the model. [3] We select 2008 as the basis year for our analysis, the year that places the respondents between the ages of 23 and 27. We make use of 5 years of data in our estimation, and 2008 also happens to be the latest year we can choose as our basis year because data are only available through 2009 at this time.

Health is measured by the respondents on a 1-5 scale, representing "excellent", "very good", "good", "fair", and "poor", respectively. We group the respondents who answered "very good" and "good" into the same category, which we call responding 2.5, because these groups are very similar and the increase in sample size allows for more accurate estimates. We would use a more objective measure of health if one was available in the NLSY97, but since the responses are all self-reported it is unclear how to create a more accurate measure of health using responses from other questions (such as height and weight).

[3] The main reason why we do not include other explanatory variables as in the drug use literature is due to the sub-sample size restriction. We do present results conditioned on gender to highlight the difference between those subpopulations.

For marijuana use, we use a variable in the data that indicates whether the respondents claim to have used marijuana at all since the date of the previous interview. To make this variable binary, we group those who admitted to using marijuana with those who did not respond to the question or said they could not remember. Participants who said they did not use marijuana in the last year make up the other group. For background information, we use the responses available from the initial 1997 survey and the 2008 survey when appropriate.

In order to run our estimation we take some steps in dropping observations. We disregard all observations (individuals) who were not interviewed for any of the 5 years our variables span. Respondents were not interviewed because they could not be tracked down or were otherwise unavailable, and since we require data on all 5 years we need to drop these people. Additionally, we drop respondents who did not respond to the health condition question. The percentage of participants who did not respond to this question in 2008 and 2007 was less than 1%. We do not believe that dropping these observations bias our results. Our sample thus reduces to 6298 individuals. Table 2.5 displays summary statistics for our sample. When conditioning on covariates, we drop observations that did not have a response for any of the variables on which we were conditioning. While an omission of a response about drug use may be informative about whether that person has actually used drugs, we do not believe the same is true for omissions on questions regarding background. Assuming that agreeing to be interviewed but not answering questions about background information is independent of true drug use, the results presented here are not biased. Tables 2.6 and 2.7 provide some reduced-form evidence on how drug use status is affected by lagged covariates, and how health status is affected by drug

use status. In those tables, all of the variables are binary except for health (H) which takes on the values discussed above. For the remaining variables, $D = 1$ if the participant used marijuana in the previous year, Post-HS Education $= 1$ if awarded any degree beyond a high school diploma by 2008, Married $= 1$ if married, White $= 1$ if white, and Male $= 1$ if male.

2.4.2 Basic Approach

We use 2008 as the basis year to do the estimation. In total there are 6298 effective observations. As is depicted in Assumption 1.5, when we do estimation for each subgroup of d_{t-1}, we would like to find the best subsample of h_{t-1} such that the eigenvalues are most distinct from each other and therefore the estimation is most valid. After trying different strategies, we decide to combine the samples where H_τ equals 2 or 3.[4] They account for nearly 70% of the total population, and this increase in the sample size for a particular subgroup facilitates more accurate estimation. Thus, the dependent variable takes the following possible values:

$$H_\tau = \begin{cases} 1 & \text{if in “excellent health condition”} \\ 2.5 & \text{if in “very good or good health condition”} \\ 4 & \text{if in “fair health condition”} \\ 5 & \text{if in “poor health condition”} \end{cases},$$

In addition, we assume,

$$E\left[H_t | D_t^* = i, d_{t-1}, H_{t-1} = 2.5\right] \neq E\left[H_t | d_{t-1}, H_{t-1} = 2.5\right], \text{for any } i.$$

which is a specification based on Assumption 1.5.

[4]For robustness check, we also provide estimation results by combining $H_\tau = 1, 2$ and 3 in the appendix.

The main estimation results are shown in Table 2.8, while Table 2.9 and Table 2.10 are the estimation results conditional on gender covariate. There are several interesting results from this table. Firstly, for any given D_{t-1}, the probability of reporting drug use for actual drug users is higher than those non-drug users. For the first two rows of each table, by assumption the truth-telling dominant rule is true; but for the third and fourth row, it still holds. For those who did not report marijuana use last year and did not use marijuana this year, the estimated probability of reporting drug use is zero. This is consistent with any intuition that leads one to believe the misreporting problem is mostly related to individuals trying to hide true drug use. The misreporting problem is most prevalent with people who did not report drug use last year and did use drug this year ($D_t^* = 1, D_{t-1} = 0$). The estimated probability of telling the truth is only 30.89%, much lower than the 82.23% where participants reported using marijuana last year. Our estimation for the true marijuana use proportion in the cohort is 32.13%, significantly higher than the reported proportion of 17.61%. These results show the significant role of misclassification error in the data. When we condition on gender and reestimate the misclassification errors, the only major difference between females and males arises when the individuals did not report using marijuana last year but use it this year (i.e. $D_t^* = 1, D_{t-1} = 0$). In that situation, males are more likely than females to hide the truth and misreport this year.

2.4.3 General Approach

In this section, we adopt the general model and estimate the misclassification matrices according to Theorem 2.2. Similar as in the basic case, we combine $H_\tau = 2$ and 3.[5] The results are presented in Tables 2.11, and results when conditioning on gender covariates are presented in Table 2.12 and Table 2.13.

Firstly, the general results are mostly consistent with those from the basic model. For instance, the people who have been using marijuana for both periods and reported using it last period have a probability of reporting usage as high as 75.02%. By contrast, people who have not used marijuana for either period and did not report usage last period have a very low probability (1.78%) of reporting usage this period. Secondly, conditional on true drug use in the previous period, people who reported using marijuana last period are more likely to report using it again this period. These results seem to display certain habits in behavior.

Another exciting result lies in the fifth row of Table 2.11, where people used marijuana last year but misreported this drug use. When continuing to use drugs this year, the probability for them to report the truth is less than 40%, much less than those who already told the truth last period (75.02%). This helps show the severity of misreporting problem in the survey.

Recall the example we used to motivate our more general model where we stressed that two individuals who were likely to have different reporting behavior would be captured by the same probability when only one latent variable was present. The comparison of the seventh and eighth row shows us the importance of incorporating last-period marijuana usage status. The rows tell us that, when

[5] In the appendix we also reestimate the measurement error matrices by combining $H_\tau = 1, 2$ and 3, as in the basic case.

people used marijuana last year but hid the truth, he or she is still much more likely to report using marijuana (24.72%) compared with those who do not use drugs for two years and do not report using it last year (1.78%).

The estimates for expected reported health conditions are more dispersed than in the basic case, as is shown in the ninth through sixteenth row of the tables. Also the standard errors are much larger compared with those in the basic case. At a price of losing the estimation accuracy of the eigenvalues, we are able to gain more accuracy of estimated eigenvectors which represent the misclassification errors of primary interest. Here the estimated eigenvalues only serve a role to distinguish different columns of eigenvectors, therefore it is acceptable in this application to have some eigenvalues that are out of proportion.

When we condition on gender covariates, the distinction between males' and females' misreporting behavior become clearer than in the basic case. Firstly consistent with the basic case, when males reported no marijuana use last period, they are more likely to follow their "habit" of saying "no" to the same question this period even though they in fact are using marijuana, as is shown by the fifth and sixth rows of Table 2.12 and Table 2.13. Moreover, when using marijuana last year and telling the truth, males are more likely than females to report "yes" to the same question this year, even when they are not using marijuana this year, (62.34% versus 39.18%). In other words, males appear to be more stuck with their last-period reporting behavior than are females, at least in some cases.

One thing to note is the probability of $\Pr(D_t = 1 | D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0)$ for males (10.11%). It is not very different from $\Pr(D_t = 1 | D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 0) = 11.16\%$ for males, which indicates that males who are non-drug

users this year and did not report using it last year have a probability as high as 10% of reporting usage this year. Somewhat counterintuitive as it seems, when we bootstrap for the mean and median, we found that they are much more distinctive from each other. The median for $\Pr(D_t = 1|D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0)$ reduces to 6.21%, which looks more reasonable. We suspect that the high point estimation is resulted from some deeper unobserved patterns within the male subgroup, which calls for further study into the misreporting problem, conditional on more covariates.

One last observation is the marginal probability of true marijuana usage. It can be seen from the last rows of Table 2.11, Table 2.12 and Table 2.13. Males have higher probability of using marijuana than do females (17.02% versus 10.43%), but all of the marginal probabilities are not quite different from the probability of reporting marijuana use. Again, if we look at the bootstrap means and medians we see results that are higher than our point estimates. This tells us that people are not always underreporting their marijuana usage, instead, there exists noisy reporting phenomenon in both directions.

2.4.4 Hypothesis Testings of the Conditional Independence Assumptions

In this section we would like to test the validity of the conditional independence assumptions we have made in deriving our identification models. Specifically

we test the following hypotheses:

2.4.4.1 Basic Approach

Firstly we would like to test whether D_{t-1} always plays a role in determining the distribution of D_t given D_t^*. Namely, the null hypothesis is:

$$H_0 : \Pr(D_t|D_t^*, D_{t-1}) = \Pr(D_t|D_t^*)$$

versus the alternative hypothesis,

$$H_1 : \Pr(D_t|D_t^*, D_{t-1}) \neq \Pr(D_t|D_t^*)$$

Here we define

$$\Lambda := \begin{bmatrix} \Pr(D_t = 1|D_t^* = 1, D_{t-1} = 1) - \Pr(D_t = 1|D_t^* = 1, D_{t-1} = 0) \\ \Pr(D_t = 1|D_t^* = 0, D_{t-1} = 1) - \Pr(D_t = 1|D_t^* = 0, D_{t-1} = 0) \end{bmatrix},$$

and test $H_0 : \Lambda(i) = 0$ versus $H_1 : \Lambda(i) \neq 0$ separately for $i = 1, 2$. Table 2.14 below displays the 95% bootstrap confidence intervals for $\Lambda(i)$. It can be indicated from the table that the Λ is significantly different from zero, therefore we could effectively reject the null hypothesis.

2.4.4.2 General Approach

Now we test the validity of the conditional independence of our general model. Similar as in the basic model, we want to test whether latent variables and the independently observed variables always play a role in determining the distribution of the proxy. Here, however, we need to do three classes of testings in total. Firstly we would like to test whether the following equality holds,

$$H_0 : \Pr(D_t|D_t^*, D_{t-1}^*, D_{t-1}) = \Pr(D_t|D_t^*, D_{t-1}^*),$$

for each given values of (D_t^*, D_{t-1}^*).

Specifically, we define Λ_1 as the L^2 norm of the difference of two probability vectors, and test whether Λ_1 is statistically significantly larger than zero. In other words,

$$\Lambda_1 := ||\Pr(D_t = 1|D_t^*, D_{t-1}^*, D_{t-1} = 1) - \Pr(D_t = 1|D_t^*, D_{t-1}^*, D_{t-1} = 1)||_{L^2},$$

where for $d = 1, 0$

$$\Pr(D_t = 1|D_t^*, D_{t-1}^*, D_{t-1} = d) := \begin{bmatrix} \Pr(D_t = 1|D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = d) \\ \Pr(D_t = 1|D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = d) \\ \Pr(D_t = 1|D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = d) \\ \Pr(D_t = 1|D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = d) \end{bmatrix}.$$

And we test

$$H_0 : \Lambda_1 = 0$$

versus

$$H_1 : \Lambda_1 > 0.$$

The second hypothesis we would like to test is

$$H_0 : \Pr(D_t|D_t^*, D_{t-1}^*, D_{t-1}) = \Pr(D_t|D_t^*, D_{t-1}),$$

for each given values of (D_t^*, D_{t-1}). Following the same logic, we define Λ_2 as the L^2 norm of the difference of two probability vectors, and test whether Λ_2 is significantly larger than zero. In other words,

$$\Lambda_2 := ||\Pr(D_t = 1|D_t^*, D_{t-1}^* = 1, D_{t-1}) - \Pr(D_t = 1|D_t^*, D_{t-1}^* = 0, D_{t-1})||_{L^2},$$

where for $d = 1, 0$

$$\Pr(D_t = 1|D_t^*, D_{t-1}^* = d, D_{t-1}) := \begin{bmatrix} \Pr(D_t = 1|D_t^* = 1, D_{t-1}^* = d, D_{t-1} = 1) \\ \Pr(D_t = 1|D_t^* = 1, D_{t-1}^* = d, D_{t-1} = 0) \\ \Pr(D_t = 1|D_t^* = 0, D_{t-1}^* = d, D_{t-1} = 1) \\ \Pr(D_t = 1|D_t^* = 0, D_{t-1}^* = d, D_{t-1} = 0) \end{bmatrix}.$$

We then test

$$H_0 : \Lambda_2 = 0$$

versus

$$H_1 : \Lambda_2 > 0.$$

Lastly we want to test whether the following equality holds

$$H_0 : \Pr(D_t | D_t^*, D_{t-1}^*, D_{t-1}) = \Pr(D_t | D_t^*),$$

for any given values of $(D_{t-1}^*,\ D_{t-1})$. To test this, we first calculate the conditional probability given any particular values of the pair $(D_{t-1}^* = d_1, D_{t-1} = d_2)$, where $d_1, d_2 \in \{0, 1\}$, and then define $\Lambda_3(d_1, d_2)$ as the L^2 norm of the difference of these two probability vectors. In other words,

$$\Lambda_3(d_1, d_2) := ||\Pr(D_t = 1 | D_t^*, D_{t-1}^* = d_1, D_{t-1} = d_2) - \Pr(D_t = 1 | D_t^*, D_{t-1}^* = d_1, D_{t-1} = d_2)||_{L^2},$$

where,

$\Pr(D_t = 1 | D_t^*, D_{t-1}^* = d_1, D_{t-1} = d_2) := \begin{bmatrix} \Pr(D_t = 1 | D_t^* = 1, D_{t-1}^* = d_1, D_{t-1} = d_2) \\ \Pr(D_t = 1 | D_t^* = 0, D_{t-1}^* = d_1, D_{t-1} = d_2) \end{bmatrix}$, for $d_1, d_2 \in \{0, 1\}$. We test whether Λ_3 is significantly larger than zero [6], i.e.

$$H_0 : \Lambda_3(d_1, d_2) = 0$$

versus

$$H_1 : \Lambda_3(d_1, d_2) > 0.$$

Table 2.15 displays the bootstrap confidence intervals for each of the statistics we use in our testings. It can be seen from the table that, none of the bootstrap confidence intervals is significantly close to zero, which means we can reject all the null hypotheses, therefore our conditional independence assumptions are valid for the general case.

[6] Note that Λ_3 is a L^2-norm which is always nonnegative, so we do one-sided test here

2.5 Conclusion

In this paper we use an eigen-decomposition method from Hu (2008) to estimate the misclassification errors in self-reported drug use. Using reasonable and relatively weak assumptions in our models, we find that the self-reported responses from the NLSY97 sometimes underreport the true level of marijuana use. Both our basic model and general model give us insights into severity and pattern of this misclassification problem. We conclude that if this misclassification problem is simply ignored in empirical work, results could be severely biased and difficult to interpret. In future work, if we can obtain more complete data detailing the different levels of marijuana use, we can attempt a continuous version of this nonparametric identification which could give us more insight in this topic.

Table 2.1: Summary of Variables - Basic Case

Variable types	Variable names	Description
Dependent variable	H_t	self-reported health condition at time t
Proxy variable	D_t	self-reported drug use status at time t
Latent variable	D_t^*	true drug use status st time t
Instrument variable	D_{t-2}	self-reported drug use status at time $t-2$
Other independently observed variables	D_{t-1}	self-reported drug use status at time $t-1$
	H_{t-1}	self-reported health condition at time $t-1$

Table 2.2: Summary of Variables - General Case

Variable types	Variable names	Description
Dependent variable	H_t	self-reported health condition at time t
Proxy variables	D_{t+1}	self-reported drug use status at time $t+1$
	D_t	self- reported drug use status at time t
Latent variables	D_t^*	true drug use status at time t
	D_{t-1}^*	true drug use status at time $t-1$
Instrument variables	D_{t-2}	self-reported drug use status at time $t-2$
	D_{t-3}	self-reported drug use status at time $t-3$
Other independently observed variables	D_{t-1}	self-reported drug use status at time $t-1$
	H_{t-1}	self-reported health condition at time $t-1$

Table 2.3: Simulation Results: Means, Medians and Standard Errors

		Sample Size				
Estimated Probabilities		7,000	10,000	50,000	100,000	True Value
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 1)$	mean	0.2073	0.2036	0.2000	0.2009	0.2000
	median	0.2040	0.2012	0.2000	0.2002	
	(std.err.)	(0.0578)	(0.0481)	(0.0212)	(0.0150)	
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 1)$	mean	0.0492	0.0482	0.0495	0.0500	0.0500
	median	0.0522	0.0501	0.0503	0.0500	
	(std.err.)	(0.0198)	(0.0169)	(0.0075)	(0.0053)	
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 0)$	mean	0.2527	0.2506	0.2499	0.2501	0.2500
	median	0.2500	0.2499	0.2498	0.2500	
	(std.err.)	(0.0289)	(0.0242)	(0.0100)	(0.0073)	
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 0)$	mean	0.0982	0.0984	0.0994	0.0999	0.1000
	median	0.0999	0.0995	0.0997	0.0999	
	(std.err.)	(0.0182)	(0.0147)	(0.0062)	(0.0042)	

Table 2.4: Simulation Results: Means, Medians and Standard Errors

		Sample size				
Estimated probabilities		7,000	10,000	50,000	100,000	True value
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 1)$	mean	0.7660	0.7779	0.8018	0.8012	0.8000
	median	0.7813	0.7859	0.8008	0.8002	
	(std.err.)	(0.1541)	(0.1399)	(0.0401)	(0.0284)	
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 1)$	mean	0.1463	0.1267	0.1006	0.0981	0.1000
	median	0.1302	0.1234	0.1006	0.0986	
	(std.err.)	(0.1288)	(0.0997)	(0.0551)	(0.0407)	
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 1)$	mean	0.5490	0.5740	0.6490	0.6790	0.7000
	median	0.5867	0.6157	0.6763	0.6932	
	(std.err.)	(0.2630)	(0.2603)	(0.1966)	(0.1397)	
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 1)$	mean	0.1271	0.1151	0.0972	0.0984	0.1000
	median	0.1130	0.1032	0.1007	0.1027	
	(std.err.)	(0.1355)	(0.1082)	(0.0575)	(0.0438)	
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 0)$	mean	0.6068	0.6076	0.6027	0.6016	0.6000
	median	0.6026	0.5987	0.5993	0.5995	
	(std.err.)	(0.1415)	(0.1136)	(0.0446)	(0.0306)	
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 0)$	mean	0.1403	0.1173	0.0988	0.0981	0.1000
	median	0.1240	0.1132	0.1023	0.1000	
	(std.err.)	(0.1293)	(0.0869)	(0.0488)	(0.0383)	
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 0)$	mean	0.4002	0.4120	0.4809	0.4936	0.5000
	median	0.4126	0.4248	0.4850	0.4994	
	(std.err.)	(0.2407)	(0.2254)	(0.1351)	(0.0947)	
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0)$	mean	0.0891	0.0724	0.0519	0.0500	0.0500
	median	0.0544	0.0520	0.0496	0.0496	
	(std.err.)	(0.1234)	(0.0820)	(0.0410)	(0.0325)	

Table 2.5: Means and Standard Deviations of Key Variables From 2008 Survey

Variable	Population	Females	Males
D_{t+1}	0.1750	0.1415	0.2108
	(0.3800)	(0.3485)	(0.4079)
D_t	0.1904	0.1645	0.2180
	(0.3926)	(0.3708)	(0.4129)
D_{t-1}	0.1851	0.1504	0.2223
	(0.3884)	(0.3575)	(0.4158)
D_{t-2}	0.1980	0.1581	0.2406
	(0.3985)	(0.3649)	(0.4275)
D_{t-3}	0.2190	0.1814	0.2590
	(0.4136)	(0.3854)	(0.4382)
H_t	2.2484	2.3264	2.1652
	(0.9428)	(0.9454)	(0.9330)
H_{t-1}	2.2188	2.3090	2.1225
	(0.9490)	(0.9497)	(0.9389)
Post-HS Education	0.3069	0.3373	0.2744
	(0.4612)	(0.4728)	(0.4463)
Married	0.2885	0.3236	0.2510
	(0.4531)	(0.4679)	(0.4337)
White	0.5087	0.4871	0.5318
	(0.5000)	(0.4999)	(0.4991)

Table 2.6: Regression Results with Covariates

	D_{t+1}	
Variable	Coefficient	Standard Error
Intercept	0.0081	0.0069
D_t	0.3466	0.0120
D_{t-1}	0.1708	0.0129
D_{t-2}	0.1582	0.0125
D_{t-3}	0.1018	0.0112
Post-HS Degree	0.0051	0.0078
Married	-0.0028	0.0079
White	0.0131	0.0073
Male	0.0173	0.0071

Table 2.7: Regression Results with Covariates

	H_t	
Variable	Coefficient	Standard Error
Intercept	1.1588	0.0323
D_{t+1}	0.0278	0.0353
D_t	0.0415	0.0358
D_{t-1}	-0.0342	0.0364
D_{t-2}	0.0224	0.0354
D_{t-3}	0.0404	0.0316
H_{t-1}	0.5326	0.0106
Post-HS Education	-0.2235	0.0221
Married	-0.0073	0.0221
White	-0.0021	0.0204
Male	-0.0842	0.0199

Table 2.8: Estimation Results: Basic Approach

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 1)$	0.8223	0.8382	0.8325	0.1200
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 1)$	0.5317	0.5157	0.5259	0.1793
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 0)$	0.3089	0.3848	0.3008	0.3067
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 0)$	0.0000	0.0328	0.0138	0.0472
$E(H_t \mid D_t^* = 1, D_{t-1} = 1)$	2.4896	2.4976	2.4932	0.0563
$E(H_t \mid D_t^* = 0, D_{t-1} = 1)$	2.3985	2.3798	2.3888	0.0761
$E(H_t \mid D_t^* = 1, D_{t-1} = 0)$	2.4419	2.4481	2.4393	0.0657
$E(H_t \mid D_t^* = 0, D_{t-1} = 0)$	2.3973	2.3936	2.3972	0.0378
$Pr(D_t^* = 1)$	0.3213	0.3409	0.2529	0.2344

Table 2.9: Estimation Results: Basic Approach for Males

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 1)$	0.7349	0.8076	0.8122	0.1652
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 1)$	0.5492	0.4179	0.4399	0.2912
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 0)$	0.1901	0.3801	0.2498	0.3502
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 0)$	0.0000	0.0376	0.0116	0.0583
$E(H_t \mid D_t^* = 1, D_{t-1} = 1)$	2.4528	2.4726	2.4650	0.0670
$E(H_t \mid D_t^* = 0, D_{t-1} = 1)$	2.4154	2.3919	2.4019	0.1064
$E(H_t \mid D_t^* = 1, D_{t-1} = 0)$	2.3770	2.3781	2.3772	0.0853
$E(H_t \mid D_t^* = 0, D_{t-1} = 0)$	2.3559	2.3508	2.3566	0.0538
$Pr(D_t^* = 1)$	0.4419	0.3945	0.3058	0.2461

Table 2.10: Estimation Results: Basic Approach for Females

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 1)$	0.8632	0.8591	0.8631	0.1270
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 1)$	0.5346	0.5143	0.5336	0.1798
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 0)$	0.3661	0.4273	0.3718	0.3165
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 0)$	0.0276	0.0357	0.0300	0.0509
$E(H_t \mid D_t^* = 1, D_{t-1} = 1)$	2.5404	2.5543	2.5459	0.1015
$E(H_t \mid D_t^* = 0, D_{t-1} = 1)$	2.3595	2.3506	2.3647	0.1211
$E(H_t \mid D_t^* = 1, D_{t-1} = 0)$	2.5271	2.5358	2.5179	0.1134
$E(H_t \mid D_t^* = 0, D_{t-1} = 0)$	2.4283	2.4178	2.4260	0.0554
$Pr(D_t^* = 1)$	0.2032	0.2993	0.1846	0.2462

Table 2.11: Estimation Results: General Approach

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 1)$	0.7502	0.7820	0.8055	0.2778
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 1)$	0.6340	0.6504	0.6415	0.2253
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 1)$	0.3022	0.4238	0.4148	0.2264
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 1)$	0.2412	0.4341	0.4326	0.2383
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 0)$	0.3956	0.4599	0.4606	0.1998
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 0)$	0.4025	0.4631	0.4543	0.1980
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 0)$	0.2472	0.2674	0.2542	0.1629
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0)$	0.0178	0.1233	0.0754	0.1653
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.1683	2.6398	2.4672	4.3786
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	2.1625	2.3381	2.4308	2.0964
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.4537	3.3446	2.5025	7.7230
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	2.4217	0.8459	2.1924	15.1518
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	2.6832	3.4901	2.6696	6.1663
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	1.1974	1.2858	2.1426	6.9415
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	2.5035	2.5830	2.4285	5.0442
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	2.3729	1.4602	2.3888	19.7184
$Pr(D_t^* = 1)$	0.1380	0.2410	0.1931	0.1656

Table 2.12: Estimation Results: General Approach for Males

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 1)$	1.0000	0.7614	0.7836	0.1767
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 1)$	0.7099	0.6190	0.6066	0.1996
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 1)$	0.6234	0.4208	0.4154	0.2210
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 1)$	0.3573	0.3723	0.3681	0.2344
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 0)$	0.2909	0.4588	0.4528	0.1964
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 0)$	0.5278	0.5189	0.5004	0.2262
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 0)$	0.1116	0.2243	0.1919	0.1454
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0)$	0.1011	0.0951	0.0621	0.1025
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.5559	2.1450	2.4369	8.9161
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	3.0644	2.5098	2.4468	7.1481
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.4689	3.9783	2.5317	23.2692
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	2.3456	1.7889	2.3204	3.4740
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	2.5426	3.1131	2.5466	10.6401
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	1.5439	0.4875	1.7965	31.7647
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	2.3243	2.7043	2.4747	7.2401
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	3.2677	2.8256	2.3581	14.2710
$Pr(D_t^* = 1)$	0.1702	0.2687	0.2370	0.1465

Table 2.13: Estimation Results: General Approach for Females

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 1)$	0.7461	0.7957	0.8258	0.1617
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 1)$	0.6377	0.6592	0.6641	0.1918
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 1)$	0.3918	0.4573	0.4644	0.2131
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 1)$	0.4031	0.4253	0.4255	0.1992
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 0)$	0.6477	0.5352	0.5361	0.2185
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 0)$	0.6527	0.4651	0.4739	0.2369
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 0)$	0.1265	0.2459	0.2186	0.1802
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0)$	0.0000	0.0963	0.0461	0.1200
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	3.7445	2.7840	2.6129	12.1329
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	1.7189	2.3308	2.5285	3.4989
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.6786	3.8359	2.5554	7.9331
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	2.3810	1.0889	2.1884	7.8487
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	5.0667	6.0379	2.7972	82.1887
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	2.5889	1.4805	2.4274	8.9413
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	2.4333	-8.6810	2.4171	423.8590
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	1.3383	2.1029	2.4237	11.8090
$Pr(D_t^* = 1)$	0.1043	0.2495	0.1737	0.2102

Table 2.14: Testing Validity of Conditional Independence – $H_t = 2, 3$ Case

Null hypothesis H_0	95% confidence interval of $\Lambda(i)$
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 0)$	[0.0000,0.9208]
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 0)$	[0.0645,0.8078]

Table 2.15: Testing Validity of Conditional Independence – $H_t = 2, 3$ Case

Null hypothesis H_0	95% confidence interval of $\Lambda_3(d_1, d_2)$
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^*, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^*, D_{t-1} = 0)$	[0.0570,1.1796]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1}) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1})$	[0.0000,0.7177]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1} = 0)$	[0.0142, 0.8735]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1} = 1)$	[0.0000, 0.5571]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1} = 0)$	[0.0105, 0.9371]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1} = 0) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1} = 1)$	[0.0070, 0.7779]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1} = 0) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1} = 0)$	[0.0000, 0.5661]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1} = 0)$	[0.0128, 0.8590]

Chapter 3

Asymmetry in the US Treasury Futures Market

3.1 Introduction

Volatility has captured the attention of those seeking to better understand risk as well as asset prices more generally, but only recently has attention turned to decomposing volatility based on the sign and magnitude of the directional move in price at high frequencies. Recent nonparametric work on modeling realized volatility stems mostly from the theoretical contributions of Barndorff-Nielsen and Shephard (2004) and Barndorff-Nielsen and Shephard (2006). In those papers, the authors discussed the convergence properties of estimators that allow researchers to separate the part of daily measured volatility that arises from continuous changes in asset prices from the volatility that is due to jumps in the price.

A growing literature has used these techniques to improve forecasting and better understand anomalies in asset prices. Andersen, Bollerslev, and Diebold (2007) looked across foreign exchange markets, equity markets, and bond yields

and found that volatility from the jump component is important, but less persistent than the volatility from the continuous component. Their results point to important improvements in volatility forecasting from decomposing volatility.

More recently, Barndorff-Nielsen, Kinnebrock, and Shephard (2010) provided additional theoretical results that allow researchers to distill the jump component from measures of positive and negative return variation, known as semivarince and more formally defined in the next section. Patton and Sheppard (2015) used these results to study the equities market and understand: i) the asymmetric effects of positive and negative returns on future volatility and ii) how these new methods can provide insight into the "leverage effect" in the equities market. As discussed in Figlewski and Wang (2000) and Glosten, Jagannathan, and Runkle (1993), among many others, the leverage effect is the idea that a drop in price of a stock will lead to higher volatility in the future (due to, perhaps, the increased leverage in the capital structure). They found negative semivariance, but not positive semivariance, to be an important predictor of future volatility in the S&P 500 Index (SPDR). Additionally, they found that the inclusion of semivariance and signed jumps can dramatically increase the quality of volatility forecasts, even beyond that which could come from an inclusion of a traditional leverage effect term. The main objective of this paper is to take techniques outlined in Barndorff-Nielsen, Kinnebrock, and Shephard (2010) and used in Patton and Sheppard (2015) to study the US Treasury futures market. Understanding how the asymmetric effects differ between the different markets could provide insight into price patterns as well as give us a deeper understanding of how these markets are linked.

The rest of the paper proceeds as follows: Section 2 describes the theoretical

results related to the semivariance and jump estimation used in our empirical analysis section. Section 3 briefly describes the data used in this analysis. Section 4 presents the results from our empirical investigation. Additionally, Section 4 extends previous work by considering measures of implied volatility when exploring the relationships between asymmetric price movements and forecasting volatility. Section 5 presents pseudo-out-of-sample estimates that compare forecasts that account for asymmetric effects from positive and negative price movements from those that do not. Section 6 concludes.

3.2 Overview of Theory and Methods

This section gives a general overview of the model and asymptotic results from Barndorff-Nielsen, Kinnebrock, and Shephard (2010) used in the empirical section. For convenience, we will attempt to use notation from Patton and Sheppard (2015) when possible, since most of our analysis serves as a comparison to the results in that paper. We start by defining the equation that describes the evolution of prices, where p_t denotes log price:

$$p_t = \int_0^t \mu_s ds + \int_0^t \sigma_s dW_s + J_t \tag{3.1}$$

In the equation above, μ_s represents the drift component, σ_t is the càdlàg volatility process, W_t is Brownian motion, and J_t is the jump process. As defined in the theoretical literature, we state the quadratic variation as:

$$[p, p] = \int_0^t \sigma_s^2 ds + \sum_{0<s\leq t} (\Delta p_s)^2 \tag{3.2}$$

where we define Δp_s to represent the jump in price, $\Delta p_s = p_s - p_{s-}$. As the

number of intervals becomes infinitely large (or equivalently the intervals themselves become infinitely small), the realized volatility converges in probability to the quadratic variation:

$$RV_t \equiv \sum_{j=1}^{m} r_{t,j}^2 \xrightarrow{p} \int_0^t \sigma_s^2 ds + \sum_{0<s\leq t} (\Delta p_s)^2 \tag{3.3}$$

where r_j is the return from one interval to the next, $r_j = p_j - p_{j-1}$, and m is the number of intervals over the selected time period. The literature has also supplied us with the convergence properties of bipower variation. This estimator is defined and converges as follows:

$$BV_t \equiv \frac{\pi}{2} \sum_{j=2}^{m} |r_{t,j}||r_{t,j-1}| \xrightarrow{p} \int_0^t \sigma_s^2 ds \tag{3.4}$$

As one can see, the BV estimator converges to only the continuous component, allowing us to do some simple arithmetic to isolate the jump:

$$RV_t - BV_t \xrightarrow{p} \sum_{0<s\leq t} (\Delta p_s)^2 \tag{3.5}$$

Barndorff-Nielsen, Kinnebrock, and Shephard (2010) proposed using positive and negative contributions to RV to say something about future predictability of total RV as well as something about the jump component. We first calculate the realized volatility associated with both positive and negative price movements (called realized semivariance):

$$RS^+ = \sum_{j=1}^{m} r_j^2 I\{r_j>0\} \tag{3.6}$$

$$RS^{-} = \sum_{j=1}^{m} r_j^2 I\{r_j<0\} \tag{3.7}$$

Barndorff-Nielsen, Kinnebrock, and Shephard (2010) showed the convergence properties of these estimators as:

$$RS^{+} = \sum_{j=1}^{m} r_j^2 I\{r_j>0\} \xrightarrow{p} \frac{1}{2}\int_0^t \sigma_s^2 ds + \sum_{0<s\leq t} \Delta p_s^2 I\{\Delta p_s>0\} \tag{3.8}$$

$$RS^{-} = \sum_{i=j}^{m} r_j^2 I\{r_j<0\} \xrightarrow{p} \frac{1}{2}\int_0^t \sigma_s^2 ds + \sum_{0<s\leq t} \Delta p_s^2 I\{\Delta p_s<0\} \tag{3.9}$$

Incorporating these components separately in a forecasting heterogenous autoregressive model (HAR) featured in Corsi (2009), instead of just the normal RV, Patton and Sheppard (2015) discovered asymmetric effects when they studied the equities market. As highlighted by other papers that have studied realized volatility across markets, the equities, Treasury, and FX market have very different historical patterns. Using these methods that incorporate potentially important asymmetries to study the Treasury market adds to our knowledge of these differences and provides further insight into the reasons for different volatility forecasts.

Using these two components, we are able to isolate signed jump variation:

$$\Delta J^2 \equiv RS^{+} - RS^{-} \xrightarrow{p} \sum_{0<s\leq t} \Delta p_s^2 I\{\Delta p_s>0\} - \sum_{0<s\leq t} \Delta p_s^2 I\{\Delta p_s<0\} \tag{3.10}$$

This measure by itself can be very insightful, but in the same way that RV can have asymmetric effects, there is reason to believe that positive and negative jumps may also have different predictive ability when looking at future

volatility. In order to add additional flexibility to our modeling, we will use the following separate measures later in our analysis:

$$\Delta J^{2+} = (RS_t^+ - RS_t^-) I\{(RS_t^+ - RS_t^-) > 0\} \tag{3.11}$$

$$\Delta J^{2-} = (RS_t^+ - RS_t^-) I\{(RS_t^+ - RS_t^-) < 0\} \tag{3.12}$$

By considering each of the elements above when forecasting RV and BV, we get a better sense of the important roles signed high frequency price moves play in forecasting future volatility over various horizons. Previous work that only considered daily signed returns, as is often found in the "leverage effect" literature, is unable to pick up on these nuances.

3.3 Data

We use high frequency 5-, 10-, and 30-year Treasury futures price data to conduct our empirical analysis. These contracts are traded on the Chicago Board of Trade (CBOT). The prices for each security are sampled at a 5-minute frequency, allowing us to distinguish relatively sudden movements from those that stretch out over the course of a day. These contracts settle quarterly, in March, June, September, and December. We handle contract rolls by setting the first 5-minute price change equal to zero, instead of the actual difference in contract prices found in the data.

We make use of data from 1990 through 2012. This span is much larger than the average span considered by most work that uses these nonparametric methods of estimating quadratic variation. Additionally, employing this span

of data allows us to study an interesting mix of pre-recession, recession, and post-recession dynamics. Lastly, we trim the sample to only include data from hours where pit trading took place at the CBOT, 8:20 ET until 15:00 ET. This is the most actively traded time for these securities, and this trimming decision follows the literature.

3.4 Empirical Analysis

When comparing results to Patton and Sheppard (2015), we will often reference the findings for the SPDR Index as those related to equities more generally. We will estimate several models, some including the asymmetric components, some including jump components, and some including both. Before considering models that predict future realized volatility, we compute some basic summary statistics and correlations between various measures used in this section and present the results in Table 3.1 and Table 3.2, respectively.

Table 3.1 looks similar to those found in the equities market in terms of their relative magnitudes among the different measures. We see that on average BV makes up a majority of RV, indicating that most of the variation in prices is driven by a smooth evolution. All significant patterns seen in the summary statistics of the data hold for all three maturities studied. Figures 3.1a-3.1c show weekly moving average plots of RV, BV, and $|\Delta J^2|$, for all three maturities.

In general, Table 3.2 is also very similar in terms of signs, magnitudes, and relative values between the different correlations, but there are some interesting distinctions both across markets and across maturities in the US Treasury futures market that are worth noting. Jumps, whether talking about signed

jumps, ΔJ^2, or the individual components, ΔJ^{2+} and ΔJ^{2-}, hold distinctly different relationships with RV and BV in the Treasuries market than in the equities market. RV and BV are much more weakly correlated and sometimes even negatively correlated with ΔJ^2. The positive correlation between these measures appears to grow as we move further along the yield curve. Although RV and BV are also negatively correlated with ΔJ^{2-} in the equites market, the negative correlation is much stronger in the Treasuries market at all maturities. Semivariances also show different relationships with other variables in these tables, and the empirical analysis below highlights some of the consequences of these different relationships.

3.4.1 HAR Model with Semivariances

Turning our attention to the linear models we employ, as a benchmark we will first estimate a HAR-RV model. We will use this to look at the relationship between past daily, weekly, and monthly RV on RV_{t+h} for $h = 1, ..., 66$ days into the future, where RV_{t+h} is the $h-$day average of RV. More formally, we will first consider:

$$\overline{RV}_{h,t+h} = \beta_0 + \beta_d RV_t + \beta_w \left(\frac{1}{4} \sum_{i=1}^{4} RV_{t-i} \right) + \beta_m \left(\frac{1}{17} \sum_{i=5}^{21} RV_{t-i} \right) + \epsilon_{t+h} \quad (3.13)$$

where $\overline{RV}_{h,t+h}$ is a simple average over the following h days:

$$\overline{RV}_{h,t+h} = \frac{1}{h} \sum_{i=1}^{h} RV_{t+i} \quad (3.14)$$

We also define the following variable:

$$\overline{RV}_{w,t} = \left(\frac{1}{4} \sum_{i=1}^{4} RV_{t-i} \right) \tag{3.15}$$

and

$$\overline{RV}_{m,t} = \left(\frac{1}{17} \sum_{i=5}^{21} RV_{t-i} \right) \tag{3.16}$$

As can been seen from the estimators, these measures represent non-overlapping weekly and monthly lags of RV. Thus our previous HAR example can be rewritten as:

$$\overline{RV}_{h,t+h} = \beta_0 + \beta_d RV_t + \beta_w \overline{RV}_{w,t} + \beta_m \overline{RV}_{m,t} + \epsilon_{t+h} \tag{3.17}$$

This equation will serve as the basis for comparison when we study the inclusion of asymmetric returns on future volatility. The results from this regression are shown in Table 3.3a - 3.3c for $h = 1$, 5, 22, and 66 on the first line for each h. Our results appear in line with those found for a similar regression on 30-year US Treasury futures in Andersen, Bollerslev, and Diebold (2007). We see a high level of persistence with $\beta_d + \beta_w + \beta_m \approx .8$, but this is weaker than that found in the equities market. Also similar to the equities market, we see the coefficient in front of daily lags, β_d, shrinks substantially as we expand the future forecasting horizon. This makes sense because today's realized volatility should intuitively have less information about the realized volatility 6 months from now than it does about tomorrow's realized volatility.

Similar to previous studies that looked at US Treasury bonds (see Andersen, Bollerslev, and Diebold (2007) and Corsi (2009)), the order of the size of the coefficients has flipped from that of the equities case. We observe that with US

Treasury futures, $\beta_d < \beta_w < \beta_m$, across all horizons and all maturities. Lastly, we find significantly smaller R^2 than those found by Patton and Sheppard (2015) in the equities market, with R^2 increasing as h increases unlike the equities market where R^2 fell after $h = 5$.

Next, we turn our attention to splitting the daily lag RV into its positive and negative components. This gives us the opportunity to see if there is added information in decomposing the RV measure, and if so, how the asymmetry could shed light on some of the volatility patters we have seen in this market. To our knowledge, we are the first to bring this type of analysis to the US Treasury market (whether futures or cash). To do so, we estimate:

$$\overline{RV}_{h,t+h} = \beta_0 + \beta_d^+ RS_t^+ + \beta_d^- RS_t^- + \beta_w \overline{RV}_{w,t} + \beta_m \overline{RV}_{m,t} + \epsilon_{t+h} \tag{3.18}$$

where the results are found on the second line for each h in Tables 3.3a - 3.3c. The general takeaway is both β_d^+ and β_d^- are positive and significant. We plot β_d, β_d^+, and β_d^- in Figures 3.2a - 3.2c for all three maturities at every h. These figures allow us to easily see that all three estimated coefficients remain positive but decline as we extend the forecasting horizon.

This may come as a surprise since Patton and Sheppard (2015) found β_d^+ to be slightly negative across all forecasting horizons and largely insignificant. Much of their later analysis used this finding to dismiss positive semivariance as they tried to determine the most useful and parsimonious forecasting model. In this sense, excluding β_d^+ from our future analysis would appear to be a mistake given our findings, and we proceed noting the relative significance of positive realized semivariance in our setting. Also worth nothing, for the 5-year futures

the magnitude of β_d^+ is greater than that of β_d^- across all horizons. As we move along to longer dated maturities we see the magnitude of β_d^- begin to exceed β_d^+ with the biggest gap between the two coming from the 30-year futures. In general, though, we find β_d^- to be substantially smaller than that found in the equities market. Taking all of this together, the impact of either positive or negative realized semivariance is more balanced in the Treasuries market than in the equities market. Although we find these coefficients to be significant, and in the case of β_d^+ the opposite sign of that found in the equities market, we see very little gain in terms of R^2 from including these terms.[1] We explore the relative benefit of using different models in our pseduo-out-of-sample analysis found at the end.

Lastly, we explicitly address the issue of the leverage effect. Patton and Sheppard (2015) discussed how past work looking to study the leverage effect has typically used an indicator function that is 1 when there is a negative daily return and interact this with lagged square return. We do something similar and interact the indicator function with daily lagged RV. Keeping the semivariance variables in the following equation allows us to say if they provide information beyond the traditional leverage effect variable. If there is no new information provided, then the coefficients in front of the two semivariances should not be statistically different and we would expect β_I to be significant, showing that the leverage effect exists.

[1] The gains cannot even be captured by our use of 3 decimal places in the table.

$$\overline{RV}_{h,t+h} = \beta_0 + \beta_d^+ RS_t^+ + \beta_d^- RS_t^- + \beta_I RV_t * I\{r_t < 0\} + \beta_w \overline{RV}_{w,t} + \beta_m \overline{RV}_{m,t} + \epsilon_{t+h} \tag{3.19}$$

The results from this regression are found on the last line for each h in Tables 3.3a - 3.3c. We find β_I to be insignificant for all maturities at all forecasting horizons when we include realized semivariances (which remain significant in most cases). Drawing the same conclusion as was found in the equities market, we find the benefit of including additional information about intraday directional price moves (semivariances) go beyond the benefits found by only considering a daily return (leverage variable). Investors make heavy use of high frequency data in modern trading strategies, and this sensitivity to intraday price development would support the findings that semivariance remains significant whereas the daily leverage effect variable is not. There appears little to be gleaned by comparing the results across maturities, as most of the changes in coefficients from including the leverage variable seem to hold broadly.

3.4.2 HAR Model with Signed Jumps and Bipower Variation

Next we consider a set of models that look to tease out the effect of jumps on future volatility. As was shown earlier, the asymptotic convergence of BV is to the continuous part of the quadratic variation, and thus we will casually think of this as representing just that. The following four equations are estimated in sequence with the results presented in Tables 3.4a - 3.4c. Equation 3.20 captures our new baseline case:

$$\overline{RV}_{h,t+h} = \beta_0 + \beta_B BV_t + \beta_w \overline{RV}_{w,t} + \beta_m \overline{RV}_{m,t} + \epsilon_{t+h} \tag{3.20}$$

The estimates continue to show similar results to those estimated from equation 3.17. Equation 3.21 captures the effect of the jump components, introduced in equations 3.11 and 3.12, on total volatility:

$$\overline{RV}_{h,t+h} = \beta_0 + \beta_J \Delta J_t^2 + \beta_B BV_t + \beta_w \overline{RV}_{w,t} + \beta_m \overline{RV}_{m,t} + \epsilon_{t+h} \tag{3.21}$$

The small and insignificant coefficients in front of the jump variable stand in contrast to the significant (and negative) coefficient found on the same variable in the equities market. This may lead one to believe that jumps are not important for forecasting volatility of Treasury futures, but this would be the wrong interpretation. Rather, it is important to consider the possibility that positive and negative jumps may affect future volatility in the same way, and therefore a coefficient that implies an increase in magnitude of positive signed jumps affects future volatility in the opposite way as an increase in magnitude of negative signed jumps would be unable to capture this nuance. The next equation adds in the proper flexibility to test this hypothesis. As we did with RV earlier, we study whether there are asymmetric effects to future volatility from the signed jump components:

$$\overline{RV}_{h,t+h} = \beta_0 + \beta_{J^+} \Delta J_t^{2+} + \beta_{J^-} \Delta J_t^{2-} + \beta_B BV_t + \beta_w \overline{RV}_{w,t} + \beta_m \overline{RV}_{m,t} + \epsilon_{t+h} \tag{3.22}$$

The results indicate that both positive and negative jumps dampen future volatility. The positive and significant coefficient in front of ΔJ_t^{2-} is the opposite of what was found for equities. We plot the coefficients for β_{J^+} and β_{J^-} in Figures 3.3a - 3.3c for all three maturities at every h. To make sure that

this finding is not due to noise in our measure of jumps, we estimate a HAR model with an added jump variable estimated according to the more traditional ratio statistic found in Tauchen and Zhou (2011). Their jump measure was not signed, and therefore our findings from estimating equation 3.22 would indicate that we should see a negative coefficient in front of the jump variable. Tables 3.5a-3.5c show these results for the three maturities where a significance threshold of .01 is used to detect jump days. The results confirm our analysis above.

There could be many reasons why both positive and negative jumps would dampen future volatility. Although we are unable to explore all the possibilities in this paper, we recognize the importance of news announcements on price discovery and jumps in the Treasuries market that may play a role. If jumps on news announcement days are associated with some resolution of uncertainty (as reflected by the jump in price of the future contract following the announcement), then this could mean less uncertainty the next day and therefore lower volatility. Further work is needed to explore this hypothesis.

Unlike equation 3.21, the following equation looks to capture the effect of the estimated jump component on future volatility coming from the continuous component:

$$\overline{BV}_{h,t+h} = \beta_0 + \beta_B BV_t + \beta_w \overline{RV}_{w,t} + \beta_m \overline{RV}_{m,t} + \epsilon_{t+h} \tag{3.23}$$

Similar to the equities case, we find that the coefficients closely match those that are estimated in equation 3.21. There is a dramatic increase in R^2, which is to be expected due to the more stable nature of that time series.

3.4.3 Inclusion of Implied Volatility

We extend the analysis presented above by bringing in a measure of implied volatility. The measure used in this paper, the MOVE Index, is produced by Bank of America Merrill Lynch. According to its description, the measure represents a "yield curve weighted index of the normalized implied volatility on 1-month Treasury options. It is the weighted average of volatilities on the CT2, CT5, CT10, and CT30."

Tables 3.6a - 3.6c display the results from the standard HAR forecasts (equations 3.17- 3.19) with the inclusion of the MOVE index as an extra covariate and Tables 3.7a - 3.7c display the results from the signed jump forecasting models (equations 3.20- 3.23) with the inclusion of the MOVE index.

The dramatic decrease in β_w and β_m by the inclusion of the MOVE index can be reconciled by the fact that implied volatility tends to be be more persistent than our nonparametric estimates of realized volatility, and therefore the more persistent weekly and monthly averages are more affected. For similar reasons the increase in R^2 should be expected. In general, significant results addressed earlier in this paper hold. We continue to see a positive coefficient on RS^+, although the significance of both RS^+ and RS^- has diminished. The relationships between the magnitude of the coefficients in front of RS^+ and RS^- continues to hold as we look across maturities. Lastly, we note that β_{J^+} and β_{J^-} remain significant at many forecasting horizons, indicating that these variables continue to provide important information regarding asymmetric effects that go beyond anything embedded in our measure of implied volatility.

3.5 Pseudo-Out-of-Sample Evidence

Lastly, we conduct some pseudo-out-of-sample analysis to compare the different forecasting models. We consider the following 4 models in this section:

$$\text{HAR-RV Model: } \hat{RV}_{h,t+h|t} = \beta_0 + \beta_d RV_t + \beta_w \overline{RV}_{w,t} + \beta_m \overline{RV}_{m,t} \tag{3.24}$$

$$\text{HAR-RS Model: } \hat{RV}_{h,t+h|t} = \beta_0 + \beta_d^+ RS_t^+ + \beta_d^- RS_t^- + \beta_w \overline{RV}_{w,t} + \beta_m \overline{RV}_{m,t} \tag{3.25}$$

$$\text{HAR-BV Model: } \hat{RV}_{h,t+h|t} = \beta_0 + \beta_B BV_t + \beta_w \overline{RV}_{w,t} + \beta_m \overline{RV}_{m,t} \tag{3.26}$$

$$\text{HAR-}\Delta J^2 \pm \text{Model: } \hat{RV}_{h,t+h|t} = \beta_0 + \beta_{J^+} \Delta J_t^{2+} + \beta_{J^-} \Delta J_t^{2-} + \beta_B BV_t + \beta_w \overline{RV}_{w,t} + \beta_m \overline{RV}_{m,t} \tag{3.27}$$

For the forecasting, we use approximately the first 10 years of data (2520 trading days) to estimate coefficients. Next, we forecast ahead the 1, 5, 22, and 66 day averages of RV. We recompute these estimates and forecasts by extending the data window one day at a time.

In order to compare the forecasts from different models, we use the methods proposed in Diebold and Mariano (1995), called the DM test statistic hereafter. To elaborate a little further, let us first define a loss function. We first consider the QLIKE loss function, defined by:[2]

$$g_1(\hat{RV}_{h,t+h|t}, \overline{RV}_{h,t+h}) = ln(\hat{RV}_{h,t+h|t}) + \frac{\overline{RV}_{h,t+h}}{\hat{RV}_{h,t+h|t}}$$

For robustness, we also consider MSE loss function:

[2]See Patton (2011) for details related to the QLIKE loss function, including the finite sample benefits when forecasting volatility.

$$g_2(e_{t+h|t}) = (e_{t+h|t})^2$$

where

$$e_{t+h|t} = \hat{RV}_{h,t+h|t} - \overline{RV}_{h,t+h}$$

represent the forecasting error. Next, we define d_t as the difference in loss functions of two different models, where those models use information from time t to forecast $\overline{RV}_{h,t+h}$:

$$d_{i,t} = g_i(\hat{RV}_{h,t+h|t}, \overline{RV}_{h,t+h}|m_1) - g_i(\hat{RV}_{h,t+h|t}, \overline{RV}_{h,t+h}|m_2), \text{ for } i = 1, 2$$

where m_1 and m_2 are added to represent different forecasting models. Under the null hypothesis $H_0 : E(d_t) = 0, \forall t$, the following is true:

$$\frac{\bar{d}}{\left(\frac{\hat{\omega}}{T}\right)^{1/2}} \to N(0, 1)$$

where $\bar{d} = \sum_{t=1}^{T} d_t$, and $\hat{\omega}$ is the consistently estimated asymptotic variance of $\bar{d}\sqrt{T}$. We use Newey-West standard errors when estimating the variance in order to control for autocorrelation. Our interpretation of the results is consistent with the approach of comparing predictive accuracy of the forecasts, not models, and thus allows us to use the standard normal limiting distribution (see Diebold (2015) for a complete discussion comparing the two interpretations). The DM test statistics are presented for the different forecasts and horizons in Tables 3.8a - 3.8c for the QLIKE loss function and Tables 3.9a - 3.9c for the

MSE loss function. A positive statistic indicates that the forecasts from asymmetric factors (the model listed second in the title of each chart) outperforms the forecasts from RV or BV (the model listed first in the title of each chart).

We can see that for nearly all horizons and across all comparisons, the forecasts generated using asymmetric components tend to outperform the forecasts generated from only lags of RV and BV. This result provides further evidence that signed high frequency returns should be included when trying to forecast volatility.

3.6 Conclusion

By considering the sign of high frequency returns, we find that negative returns that contribute to the our realized volatility estimator, negative semivariance, tend to have stronger effects on future volatility than positive returns. Unlike the results found for the SPDR index in Patton and Sheppard (2015), though, we do find significant coefficients in front of our positive realized semivariance. An interesting pattern emerges between the magnitudes of coefficients in front of positive and negative semivariance when looking over different maturities, with the magnitude of the negative semivariance coefficient dominating the positive semivarance as one moves from shorter to longer maturities.

When considering volatility forecasts using our estimated signed jump component, we find both positive and negative jumps dampen future volatility, which is also different than that found in the equities market. It is beyond the scope of this paper to conclude why these differences exist, but if one believes that jumps are associated with the market digesting new information, it would

be worth investigating whether the common prescheduled news announcements related to macro fundamentals play a role in these discovered differences in coefficients.

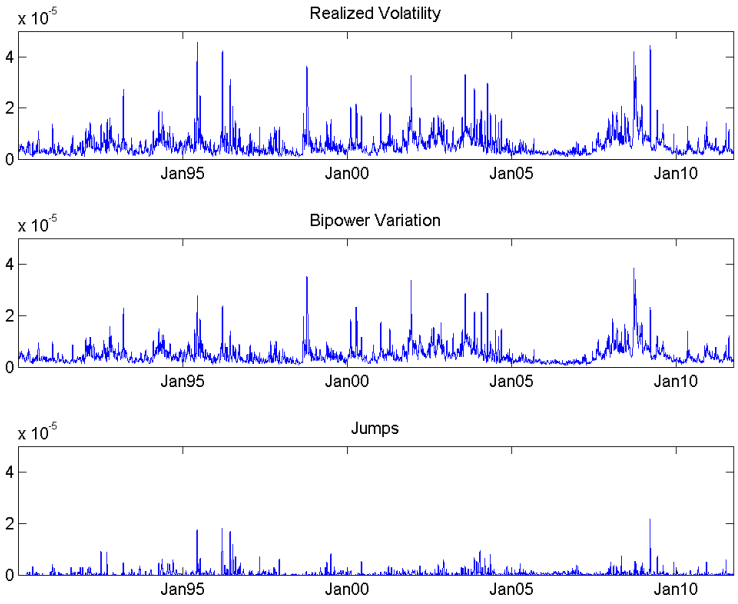


Figure 3.1a: Weekly Moving Averages for 5-Year Contract

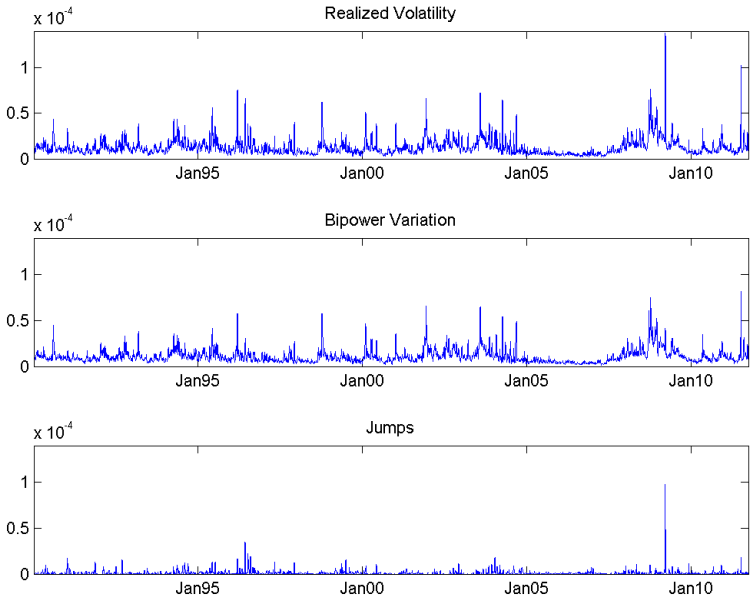


Figure 3.1b: Weekly Moving Averages for 10-Year Contract

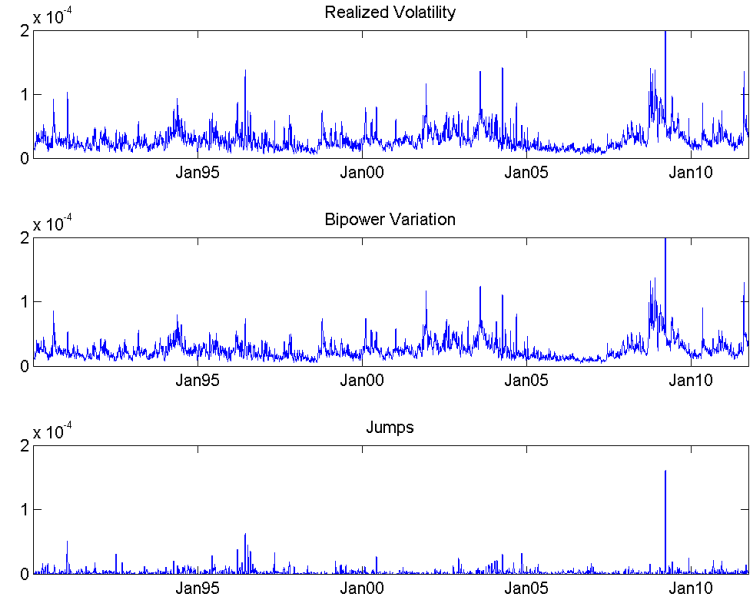


Figure 3.1c: Weekly Moving Averages for 30-Year Contract

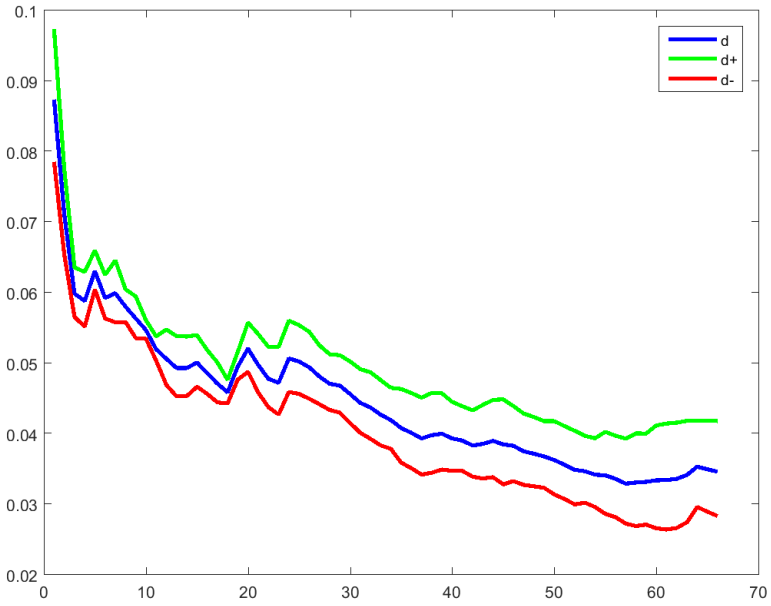


Figure 3.2a: Coefficients on RV and Signed RS for 5-Year Futures Contract

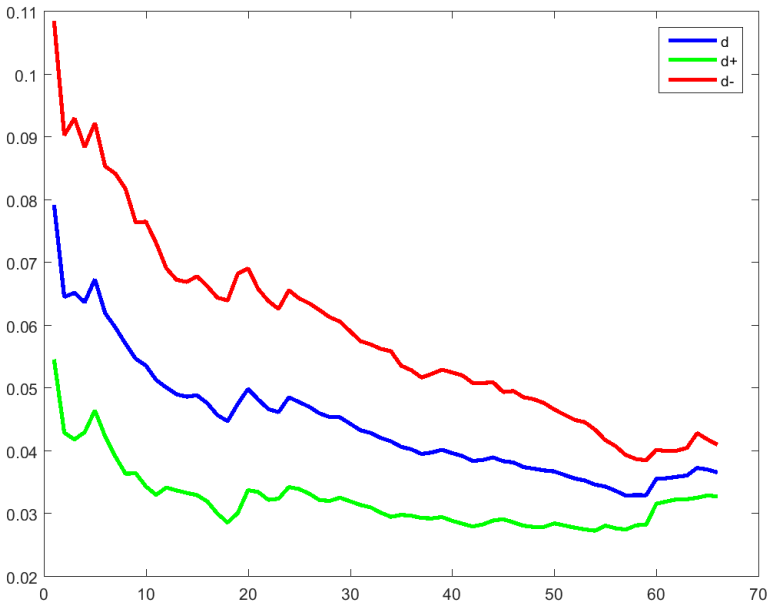


Figure 3.2b: Coefficients on RV and Signed RS for 10-Year Futures Contract

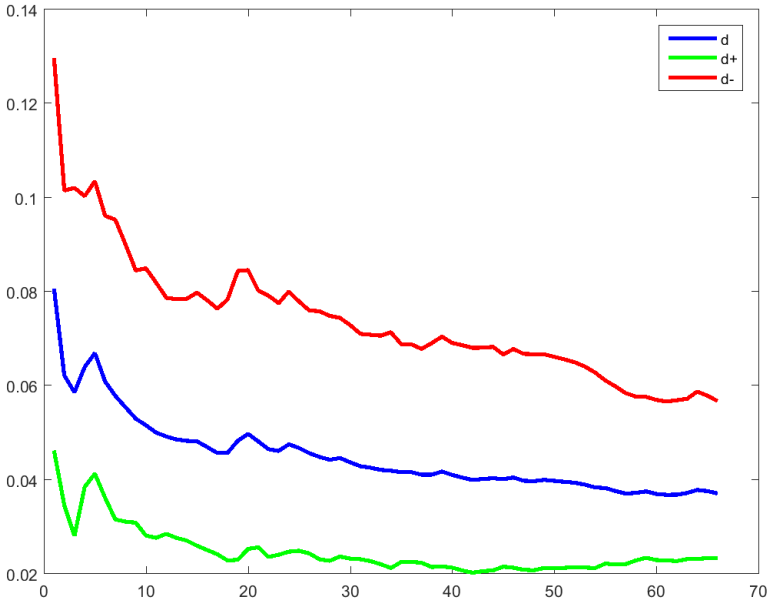


Figure 3.2c: Coefficients on RV and Signed RS for 30-Year Futures Contract

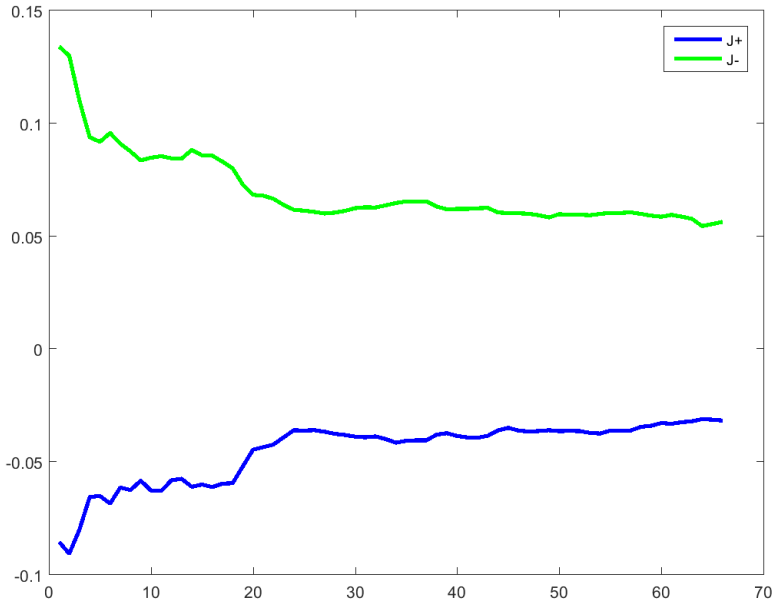


Figure 3.3a: Coefficients on Signed ΔJ^2 for 5-Year Futures Contract

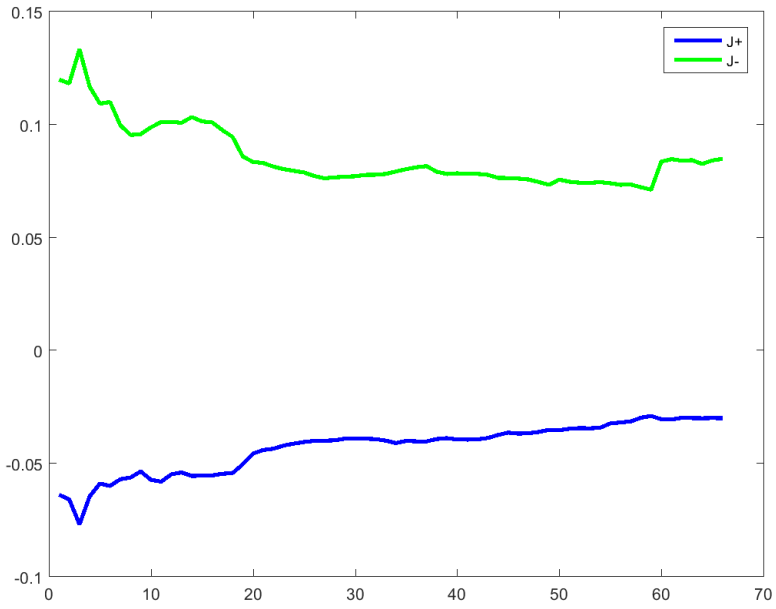


Figure 3.3b: Coefficients on Signed ΔJ^2 for 10-Year Futures Contract

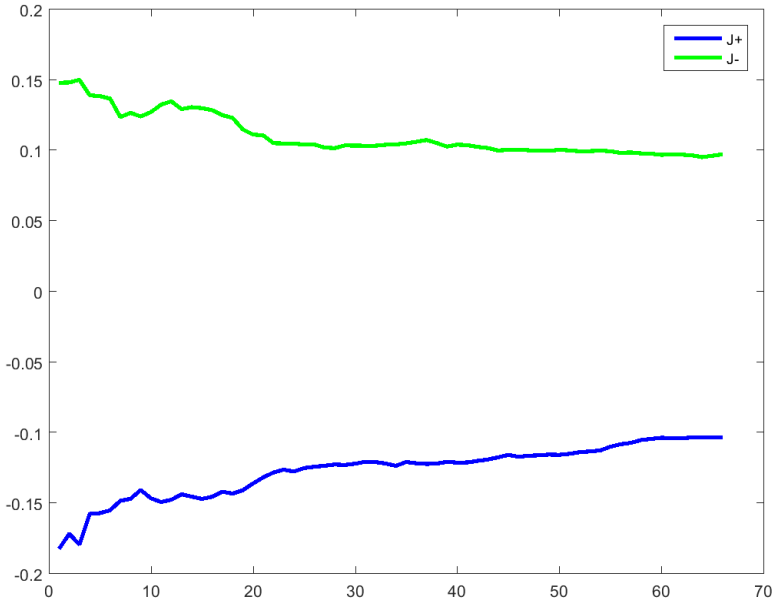


Figure 3.3c: Coefficients on Signed ΔJ^2 for 30-Year Futures Contract

Table 3.1: Summary Statistics for All Maturities

		Mean	St. Dev.	$Q_{.25}$	Median	$Q_{.75}$
5-year	RV	0.057	0.085	0.021	0.035	0.063
	BV	0.049	0.063	0.019	0.032	0.056
	RS^+	0.028	0.050	0.010	0.016	0.031
	RS^-	0.029	0.054	0.009	0.016	0.030
	ΔJ	-0.001	0.060	-0.007	-0.000	0.007
	ΔJ^+	0.009	0.040	0.000	0.000	0.007
	ΔJ^-	-0.010	0.043	-0.007	-0.000	0.000
10-year	RV	0.129	0.176	0.055	0.088	0.149
	BV	0.112	0.127	0.049	0.079	0.132
	RS^+	0.064	0.114	0.026	0.042	0.072
	RS^-	0.065	0.102	0.025	0.041	0.072
	ΔJ	-0.001	0.125	-0.016	-0.000	0.015
	ΔJ^+	0.019	0.095	0.000	0.000	0.015
	ΔJ^-	-0.020	0.077	-0.016	-0.000	0.000
30-year	RV	0.296	0.363	0.140	0.217	0.349
	BV	0.258	0.263	0.122	0.192	0.311
	RS^+	0.145	0.242	0.065	0.104	0.168
	RS^-	0.151	0.194	0.064	0.102	0.173
	ΔJ	-0.006	0.247	-0.041	-0.001	0.036
	ΔJ^+	0.039	0.188	0.000	0.000	0.036
	ΔJ^-	-0.045	0.149	-0.041	-0.001	0.000

Note: All values are scaled by 1000.

Table 3.2: Correlation Between Measures for All Maturities

		RV	BV	RS^+	RS^-	ΔJ	ΔJ^+	ΔJ^-
5-year	RV	1.000	0.918	0.800	0.832	-0.081	0.571	-0.650
	BV	0.918	1.000	0.696	0.799	-0.138	0.381	-0.552
	RS^+	0.800	0.696	1.000	0.333	0.533	0.909	-0.103
	RS^-	0.832	0.799	0.333	1.000	-0.621	0.057	-0.927
	ΔJ	-0.081	-0.138	0.533	-0.621	1.000	0.704	0.746
	ΔJ^+	0.571	0.381	0.909	0.057	0.704	1.000	0.053
	ΔJ^-	-0.650	-0.552	-0.103	-0.927	0.746	0.053	1.000
10-year	RV	1.000	0.886	0.836	0.792	0.112	0.619	-0.582
	BV	0.886	1.000	0.624	0.832	-0.113	0.300	-0.553
	RS^+	0.836	0.624	1.000	0.328	0.639	0.915	-0.090
	RS^-	0.792	0.832	0.328	1.000	-0.517	0.048	-0.902
	ΔJ	0.112	-0.113	0.639	-0.517	1.000	0.790	0.653
	ΔJ^+	0.619	0.300	0.915	0.048	0.790	1.000	0.051
	ΔJ^-	-0.582	-0.553	-0.090	-0.902	0.653	0.051	1.000
30-year	RV	1.000	0.932	0.868	0.786	0.233	0.708	-0.504
	BV	0.932	1.000	0.768	0.785	0.136	0.540	-0.455
	RS^+	0.868	0.768	1.000	0.376	0.685	0.938	-0.045
	RS^-	0.786	0.785	0.376	1.000	-0.418	0.154	-0.886
	ΔJ	0.233	0.136	0.685	-0.418	1.000	0.798	0.652
	ΔJ^+	0.708	0.540	0.938	0.154	0.798	1.000	0.063
	ΔJ^-	-0.504	-0.455	-0.045	-0.886	0.652	0.063	1.000

Table 3.3a: HAR Estimation for 5-Year

	β_d	β_d^+	β_d^-	β_I	β_w	β_m	R^2
$h = 1$	0.087				0.202	0.405	0.080
	(0.020)				(0.031)	(0.045)	
		0.097	0.079		0.203	0.405	0.080
		(0.036)	(0.031)		(0.031)	(0.045)	
		0.098	0.076	0.002	0.203	0.405	0.080
		(0.036)	(0.076)	(0.053)	(0.031)	(0.045)	
$h = 5$	0.063				0.191	0.426	0.240
	(0.013)				(0.030)	(0.043)	
		0.066	0.060		0.191	0.426	0.240
		(0.020)	(0.020)		(0.030)	(0.043)	
		0.064	0.074	-0.011	0.190	0.426	0.240
		(0.021)	(0.040)	(0.028)	(0.030)	(0.043)	
$h = 22$	0.048				0.169	0.360	0.342
	(0.011)				(0.027)	(0.039)	
		0.052	0.044		0.169	0.360	0.342
		(0.018)	(0.015)		(0.027)	(0.039)	
		0.050	0.060	-0.013	0.169	0.360	0.342
		(0.018)	(0.027)	(0.016)	(0.027)	(0.039)	
$h = 66$	0.035				0.119	0.315	0.325
	(0.006)				(0.016)	(0.044)	
		0.042	0.028		0.119	0.315	0.325
		(0.012)	(0.008)		(0.016)	(0.044)	
		0.038	0.050	-0.018	0.119	0.315	0.325
		(0.013)	(0.025)	(0.020)	(0.016)	(0.044)	

Note: The first line for each value of h corresponds to the results from estimating equation 3.17, the second line corresponds to equation 3.18, the third line corresponds to equation 3.19. Newey-West standard errors are in parentheses with bandwidth selections of 5 ($h = 1$), 8 ($h = 5$), 42 ($h = 22$), and 130 ($h = 66$).

Table 3.3b: HAR Estimation for 10-Year

	β_d	β_d^+	β_d^-	β_I	β_w	β_m	R^2
$h = 1$	0.079				0.227	0.394	0.082
	(0.024)				(0.046)	(0.057)	
		0.055	0.109		0.226	0.393	0.082
		(0.033)	(0.039)		(0.047)	(0.056)	
		0.061	0.071	0.031	0.227	0.393	0.083
		(0.035)	(0.068)	(0.046)	(0.047)	(0.056)	
$h = 5$	0.067				0.185	0.428	0.241
	(0.017)				(0.037)	(0.056)	
		0.047	0.092		0.184	0.427	0.242
		(0.026)	(0.027)		(0.038)	(0.055)	
		0.045	0.101	-0.007	0.184	0.427	0.242
		(0.029)	(0.056)	(0.035)	(0.038)	(0.055)	
$h = 22$	0.047				0.161	0.389	0.363
	(0.011)				(0.030)	(0.047)	
		0.032	0.064		0.161	0.388	0.363
		(0.020)	(0.021)		(0.030)	(0.047)	
		0.032	0.063	0.001	0.161	0.388	0.363
		(0.022)	(0.039)	(0.022)	(0.031)	(0.047)	
$h = 66$	0.037				0.126	0.347	0.363
	(0.008)				(0.022)	(0.053)	
		0.033	0.041		0.126	0.347	0.363
		(0.017)	(0.016)		(0.022)	(0.053)	
		0.031	0.055	-0.011	0.126	0.347	0.363
		(0.018)	(0.033)	(0.018)	(0.022)	(0.053)	

Note: The first line for each value of h corresponds to the results from estimating equation 3.17, the second line corresponds to equation 3.18, the third line corresponds to equation 3.19. Newey-West standard errors are in parentheses with bandwidth selections of 5 ($h = 1$), 8 ($h = 5$), 42 ($h = 22$), and 130 ($h = 66$).

Table 3.3c: HAR Estimation for 30-Year

	β_d	β_d^+	β_d^-	β_I	β_w	β_m	R^2
$h = 1$	0.081				0.223	0.431	0.097
	(0.031)				(0.063)	(0.093)	
		0.046	0.130		0.222	0.427	0.098
		(0.048)	(0.045)		(0.063)	(0.089)	
		0.064	0.047	0.064	0.222	0.429	0.099
		(0.043)	(0.067)	(0.046)	(0.064)	(0.090)	
$h = 5$	0.067				0.181	0.470	0.281
	(0.025)				(0.055)	(0.096)	
		0.041	0.103		0.180	0.467	0.282
		(0.036)	(0.032)		(0.056)	(0.094)	
		0.038	0.118	-0.012	0.179	0.467	0.282
		(0.041)	(0.062)	(0.036)	(0.056)	(0.093)	
$h = 22$	0.047				0.160	0.456	0.438
	(0.015)				(0.042)	(0.079)	
		0.024	0.079		0.159	0.454	0.440
		(0.026)	(0.030)		(0.043)	(0.079)	
		0.025	0.074	0.004	0.159	0.454	0.440
		(0.029)	(0.044)	(0.018)	(0.043)	(0.079)	
$h = 66$	0.037				0.131	0.410	0.441
	(0.009)				(0.026)	(0.062)	
		0.023	0.057		0.131	0.408	0.442
		(0.019)	(0.025)		(0.027)	(0.062)	
		0.024	0.052	0.004	0.131	0.408	0.442
		(0.022)	(0.039)	(0.016)	(0.027)	(0.062)	

Note: The first line for each value of h corresponds to the results from estimating equation 3.17, the second line corresponds to equation 3.18, the third line corresponds to equation 3.19. Newey-West standard errors are in parentheses with bandwidth selections of 5 ($h = 1$), 8 ($h = 5$), 42 ($h = 22$), and 130 ($h = 66$).

Table 3.4a: HAR Estimation with Signed Jumps for 5-Year

	$Measure$	β_J	β_J^+	β_J^-	β_B	β_w	β_m	R^2
$h = 1$	RV				0.146	0.187	0.388	0.083
					(0.030)	(0.029)	(0.045)	
	RV	0.021			0.149	0.187	0.387	0.083
		(0.025)			(0.029)	(0.029)	(0.045)	
	RV		-0.085	0.134	0.223	0.170	0.375	0.086
			(0.050)	(0.034)	(0.046)	(0.028)	(0.045)	
	BV	0.015			0.159	0.181	0.327	0.132
		(0.025)			(0.028)	(0.027)	(0.038)	
$h = 5$	RV				0.105	0.180	0.414	0.246
					(0.022)	(0.028)	(0.043)	
	RV	0.011			0.107	0.179	0.413	0.246
		(0.013)			(0.021)	(0.028)	(0.043)	
	RV		-0.065	0.092	0.159	0.167	0.405	0.251
			(0.025)	(0.023)	(0.032)	(0.028)	(0.043)	
	BV	0.010			0.110	0.163	0.357	0.300
		(0.013)			(0.020)	(0.025)	(0.037)	
$h = 22$	RV				0.079	0.161	0.351	0.348
					(0.019)	(0.025)	(0.040)	
	RV	0.010			0.081	0.160	0.350	0.349
		(0.011)			(0.018)	(0.025)	(0.040)	
	RV		-0.042	0.067	0.117	0.152	0.345	0.353
			(0.023)	(0.021)	(0.029)	(0.024)	(0.041)	
	BV	0.013			0.079	0.142	0.308	0.366
		(0.011)			(0.017)	(0.023)	(0.032)	
$h = 66$	RV				0.058	0.113	0.308	0.330
					(0.011)	(0.015)	(0.044)	
	RV	0.011			0.060	0.112	0.307	0.331
		(0.007)			(0.011)	(0.015)	(0.043)	
	RV		-0.032	0.056	0.090	0.106	0.303	0.335
			(0.016)	(0.016)	(0.019)	(0.014)	(0.043)	
	BV	0.013			0.058	0.100	0.260	0.317
		(0.008)			(0.011)	(0.014)	(0.039)	

Note: The first line for each value of h corresponds to the results from estimating equation 3.20, the second line corresponds to equation 3.21, the third line corresponds to 3.22, and the fourth from equation 3.23. Newey-West standard errors are in parentheses with bandwidth selections of 5 ($h = 1$), 8 ($h = 5$), 42 ($h = 22$), and 130 ($h = 66$).

Table 3.4b: HAR Estimation with Signed Jumps for 10-Year

	$Measure$	β_J	β_J^+	β_J^-	β_B	β_w	β_m	R^2
$h=1$	RV				0.152	0.209	0.371	0.087
					(0.035)	(0.045)	(0.053)	
	RV	0.003			0.152	0.209	0.371	0.087
		(0.022)			(0.034)	(0.045)	(0.053)	
	RV		-0.064	0.120	0.212	0.195	0.362	0.089
			(0.024)	(0.043)	(0.048)	(0.046)	(0.052)	
	BV	0.005			0.161	0.195	0.316	0.145
		(0.020)			(0.033)	(0.041)	(0.040)	
$h=5$	RV				0.127	0.170	0.409	0.251
					(0.023)	(0.036)	(0.053)	
	RV	0.002			0.127	0.170	0.409	0.251
		(0.016)			(0.022)	(0.036)	(0.053)	
	RV		-0.059	0.109	0.181	0.158	0.401	0.257
			(0.016)	(0.033)	(0.030)	(0.035)	(0.051)	
	BV	0.001			0.126	0.160	0.351	0.316
		(0.015)			(0.021)	(0.035)	(0.041)	
$h=22$	RV				0.090	0.150	0.375	0.373
					(0.019)	(0.030)	(0.046)	
	RV	0.002			0.091	0.150	0.375	0.373
		(0.013)			(0.018)	(0.029)	(0.046)	
	RV		-0.043	0.081	0.131	0.141	0.369	0.379
			(0.009)	(0.029)	(0.028)	(0.029)	(0.045)	
	BV	0.005			0.089	0.137	0.318	0.381
		(0.013)			(0.017)	(0.028)	(0.040)	
$h=66$	RV				0.072	0.117	0.335	0.371
					(0.017)	(0.021)	(0.050)	
	RV	0.011			0.073	0.117	0.335	0.372
		(0.010)			(0.016)	(0.020)	(0.050)	
	RV		-0.030	0.085	0.114	0.108	0.327	0.379
			(0.007)	(0.025)	(0.024)	(0.019)	(0.048)	
	BV	0.012			0.069	0.103	0.268	0.332
		(0.011)			(0.014)	(0.020)	(0.044)	

Note: The first line for each value of h corresponds to the results from estimating equation 3.20, the second line corresponds to equation 3.21, the third line corresponds to 3.22, and the fourth from equation 3.23. Newey-West standard errors are in parentheses with bandwidth selections of 5 ($h=1$), 8 ($h=5$), 42 ($h=22$), and 130 ($h=66$).

Table 3.4c: HAR Estimation with Signed Jumps for 30-Year

	$Measure$	β_J	β_J^+	β_J^-	β_B	β_w	β_m	R^2
$h=1$	RV				0.161	0.203	0.400	0.103
					(0.041)	(0.056)	(0.091)	
	RV	-0.036			0.167	0.202	0.397	0.103
		(0.045)			(0.033)	(0.057)	(0.086)	
	RV		-0.182	0.148	0.283	0.178	0.372	0.108
			(0.047)	(0.036)	(0.039)	(0.054)	(0.078)	
	BV	-0.030			0.175	0.195	0.350	0.178
		(0.042)			(0.033)	(0.058)	(0.069)	
$h=5$	RV				0.131	0.165	0.445	0.291
					(0.031)	(0.050)	(0.093)	
	RV	-0.026			0.135	0.164	0.443	0.292
		(0.033)			(0.026)	(0.051)	(0.091)	
	RV		-0.157	0.139	0.239	0.142	0.420	0.304
			(0.042)	(0.039)	(0.036)	(0.049)	(0.083)	
	BV	-0.026			0.136	0.159	0.396	0.377
		(0.033)			(0.026)	(0.050)	(0.073)	
$h=22$	RV				0.095	0.148	0.437	0.450
					(0.018)	(0.039)	(0.077)	
	RV	-0.024			0.099	0.147	0.435	0.451
		(0.029)			(0.018)	(0.041)	(0.077)	
	RV		-0.128	0.105	0.181	0.130	0.417	0.465
			(0.047)	(0.034)	(0.042)	(0.041)	(0.076)	
	BV	-0.021			0.098	0.140	0.387	0.469
		(0.028)			(0.017)	(0.039)	(0.070)	
$h=66$	RV				0.077	0.122	0.393	0.452
					(0.014)	(0.024)	(0.059)	
	RV	-0.014			0.079	0.122	0.392	0.452
		(0.023)			(0.016)	(0.025)	(0.058)	
	RV		-0.103	0.097	0.150	0.107	0.376	0.466
			(0.038)	(0.031)	(0.038)	(0.024)	(0.055)	
	BV	-0.012			0.076	0.113	0.340	0.424
		(0.022)			(0.015)	(0.025)	(0.052)	

Note: The first line for each value of h corresponds to the results from estimating equation 3.20, the second line corresponds to equation 3.21, the third line corresponds to 3.22, and the fourth from equation 3.23. Newey-West standard errors are in parentheses with bandwidth selections of 5 ($h=1$), 8 ($h=5$), 42 ($h=22$), and 130 ($h=66$).

Table 3.5a: HAR Estimation for 5-Year with Alternative Definition of Jump

	β_d	β_J	β_w	β_m	R^2
$h = 1$	0.168	-0.243	0.179	0.383	0.134
	(0.037)	(0.078)	(0.029)	(0.045)	
$h = 5$	0.124	-0.184	0.173	0.409	0.304
	(0.027)	(0.048)	(0.028)	(0.043)	
$h = 22$	0.095	-0.141	0.155	0.347	0.370
	(0.024)	(0.044)	(0.024)	(0.041)	
$h = 66$	0.070	-0.106	0.109	0.305	0.320
	(0.016)	(0.031)	(0.014)	(0.044)	

Note: Jumps are defined according to a ratio-statistic method (see Tauchen and Zhou (2011) equation (8)), where we set the significance level to be $\alpha = .01$. Newey-West standard errors are in parentheses with bandwidth selections of 5 ($h = 1$), 8 ($h = 5$), 42 ($h = 22$), and 130 ($h = 66$).

Table 3.5b: HAR Estimation for 10-Year with Alternative Definition of Jump

	β_d	β_J	β_w	β_m	R^2
$h = 1$	0.158	-0.214	0.205	0.369	0.145
	(0.037)	(0.054)	(0.045)	(0.053)	
$h = 5$	0.129	-0.168	0.168	0.408	0.317
	(0.025)	(0.035)	(0.036)	(0.053)	
$h = 22$	0.096	-0.133	0.147	0.373	0.384
	(0.022)	(0.030)	(0.029)	(0.046)	
$h = 66$	0.075	-0.099	0.116	0.334	0.332
	(0.020)	(0.028)	(0.020)	(0.050)	

Note: Jumps are defined according to a ratio-statistic method (see Tauchen and Zhou (2011) equation (8)), where we set the significance level to be $\alpha = .01$. Newey-West standard errors are in parentheses with bandwidth selections of 5 ($h = 1$), 8 ($h = 5$), 42 ($h = 22$), and 130 ($h = 66$).

Table 3.5c: HAR Estimation for 30-Year with Alternative Definition of Jump

	β_d	β_J	β_w	β_m	R^2
$h = 1$	0.232	-0.436	0.184	0.372	0.188
	(0.035)	(0.081)	(0.054)	(0.077)	
$h = 5$	0.173	-0.306	0.153	0.428	0.387
	(0.026)	(0.062)	(0.050)	(0.086)	
$h = 22$	0.135	-0.256	0.137	0.421	0.482
	(0.029)	(0.073)	(0.041)	(0.077)	
$h = 66$	0.108	-0.202	0.114	0.381	0.435
	(0.027)	(0.063)	(0.025)	(0.056)	

Note: Jumps are defined according to a ratio-statistic method (see Tauchen and Zhou (2011) equation (8)), where we set the significance level to be $\alpha = .01$. Newey-West standard errors are in parentheses with bandwidth selections of 5 ($h = 1$), 8 ($h = 5$), 42 ($h = 22$), and 130 ($h = 66$).

Table 3.6a: HAR Estimation with Implied Volatility for 5-Year

	β_d	β_d^+	β_d^-	β_I	β_{IV}	β_w	β_m	R^2
$h=1$	0.051				0.965	0.084	0.009	0.110
	(0.018)				(0.109)	(0.030)	(0.055)	
		0.067	0.037		0.966	0.083	0.008	0.110
		(0.032)	(0.027)		(0.109)	(0.030)	(0.055)	
		0.070	0.016	0.017	0.967	0.083	0.008	0.110
		(0.032)	(0.074)	(0.053)	(0.109)	(0.030)	(0.054)	
$h=5$	0.035				0.737	0.100	0.123	0.298
	(0.012)				(0.085)	(0.030)	(0.050)	
		0.043	0.029		0.738	0.100	0.122	0.298
		(0.018)	(0.016)		(0.085)	(0.030)	(0.050)	
		0.043	0.028	0.001	0.738	0.100	0.122	0.298
		(0.019)	(0.036)	(0.027)	(0.085)	(0.030)	(0.050)	
$h=22$	0.027				0.538	0.102	0.139	0.403
	(0.011)				(0.092)	(0.028)	(0.058)	
		0.035	0.020		0.539	0.102	0.138	0.403
		(0.017)	(0.012)		(0.092)	(0.028)	(0.058)	
		0.035	0.026	-0.005	0.539	0.102	0.138	0.403
		(0.017)	(0.023)	(0.016)	(0.091)	(0.028)	(0.058)	
$h=66$	0.022				0.340	0.077	0.175	0.360
	(0.007)				(0.094)	(0.019)	(0.060)	
		0.031	0.014		0.341	0.077	0.174	0.361
		(0.011)	(0.007)		(0.094)	(0.019)	(0.060)	
		0.028	0.029	-0.013	0.340	0.077	0.175	0.361
		(0.012)	(0.023)	(0.020)	(0.094)	(0.019)	(0.059)	

Note: This table uses the same equations as those outlined in Tables 3.3a-3.3c with the addition of the MOVE index. We scale to move index by dividing it by 10^7 in order to have it on the same order of magnitude as RV. Newey-West standard errors are in parentheses with bandwidth selections of 5 ($h=1$), 8 ($h=5$), 42 ($h=22$), and 130 ($h=66$).

Table 3.6b: HAR Estimation with Implied Volatility for 10-Year

	β_d	β_d^+	β_d^-	β_I	β_{IV}	β_w	β_m	R^2
$h = 1$	0.036				2.236	0.085	-0.045	0.118
	(0.017)				(0.206)	(0.038)	(0.043)	
		0.029	0.044		2.232	0.085	-0.044	0.118
		(0.023)	(0.030)		(0.203)	(0.038)	(0.043)	
		0.033	0.016	0.023	2.230	0.086	-0.044	0.118
		(0.023)	(0.058)	(0.042)	(0.203)	(0.038)	(0.043)	
$h = 5$	0.031				1.861	0.067	0.063	0.322
	(0.013)				(0.179)	(0.026)	(0.042)	
		0.025	0.039		1.857	0.067	0.063	0.322
		(0.018)	(0.020)		(0.177)	(0.026)	(0.042)	
		0.022	0.056	-0.014	1.857	0.067	0.063	0.322
		(0.019)	(0.045)	(0.031)	(0.177)	(0.026)	(0.042)	
$h = 22$	0.017				1.510	0.065	0.092	0.467
	(0.009)				(0.228)	(0.026)	(0.053)	
		0.015	0.020		1.508	0.065	0.092	0.468
		(0.015)	(0.012)		(0.224)	(0.026)	(0.053)	
		0.014	0.025	-0.004	1.508	0.065	0.092	0.468
		(0.017)	(0.021)	(0.016)	(0.224)	(0.026)	(0.053)	
$h = 66$	0.014				1.118	0.052	0.120	0.438
	(0.007)				(0.199)	(0.018)	(0.044)	
		0.018	0.009		1.120	0.052	0.120	0.438
		(0.012)	(0.008)		(0.199)	(0.018)	(0.045)	
		0.017	0.014	-0.004	1.120	0.052	0.120	0.438
		(0.013)	(0.019)	(0.015)	(0.199)	(0.018)	(0.045)	

Note: This table uses the same equations as those outlined in Tables 3.3a-3.3c with the addition of the MOVE index. We scale to move index by dividing it by 10^7 in order to have it on the same order of magnitude as RV. Newey-West standard errors are in parentheses with bandwidth selections of 5 ($h = 1$), 8 ($h = 5$), 42 ($h = 22$), and 130 ($h = 66$).

Table 3.6c: HAR Estimation with Implied Volatility for 30-Year

	β_d	β_d^+	β_d^-	β_I	β_{IV}	β_w	β_m	R^2
$h = 1$	0.049				3.499	0.126	0.141	0.123
	(0.020)				(0.336)	(0.037)	(0.062)	
		0.034	0.071		3.473	0.127	0.142	0.123
		(0.033)	(0.036)		(0.331)	(0.037)	(0.062)	
		0.051	-0.005	0.059	3.468	0.127	0.143	0.124
		(0.028)	(0.061)	(0.046)	(0.330)	(0.038)	(0.063)	
$h = 5$	0.040				2.995	0.098	0.221	0.340
	(0.017)				(0.355)	(0.032)	(0.069)	
		0.031	0.054		2.979	0.098	0.222	0.340
		(0.024)	(0.023)		(0.348)	(0.032)	(0.070)	
		0.027	0.074	-0.016	2.981	0.097	0.221	0.340
		(0.029)	(0.052)	(0.034)	(0.348)	(0.031)	(0.069)	
$h = 22$	0.024				2.521	0.090	0.243	0.515
	(0.013)				(0.594)	(0.036)	(0.072)	
		0.015	0.037		2.505	0.090	0.243	0.515
		(0.019)	(0.016)		(0.578)	(0.036)	(0.072)	
		0.015	0.037	0.000	2.505	0.090	0.243	0.515
		(0.021)	(0.026)	(0.015)	(0.579)	(0.036)	(0.072)	
$h = 66$	0.018				2.138	0.071	0.227	0.515
	(0.007)				(0.559)	(0.020)	(0.044)	
		0.016	0.021		2.135	0.071	0.227	0.515
		(0.013)	(0.012)		(0.554)	(0.020)	(0.045)	
		0.016	0.019	0.002	2.134	0.071	0.227	0.515
		(0.015)	(0.021)	(0.014)	(0.554)	(0.020)	(0.045)	

Note: This table uses the same equations as those outlined in Tables 3.3a-3.3c with the addition of the MOVE index. We scale to move index by dividing it by 10^7 in order to have it on the same order of magnitude as RV. Newey-West standard errors are in parentheses with bandwidth selections of 5 ($h = 1$), 8 ($h = 5$), 42 ($h = 22$), and 130 ($h = 66$).

Table 3.7a: HAR Estimation with Signed Jumps and Implied Volatility for 5-Year

	$Measure$	β_J	β_J^+	β_J^-	β_{IV}	β_B	β_w	β_m	R^2
$h=1$	RV				0.940	0.089	0.077	0.008	0.112
					(0.109)	(0.028)	(0.029)	(0.054)	
	RV	0.022			0.941	0.092	0.077	0.007	0.112
		(0.022)			(0.109)	(0.027)	(0.029)	(0.054)	
	RV		-0.059	0.109	0.924	0.150	0.066	0.005	0.113
			(0.049)	(0.033)	(0.110)	(0.044)	(0.028)	(0.054)	
	BV	0.016			0.756	0.113	0.093	0.021	0.166
		(0.022)			(0.086)	(0.026)	(0.025)	(0.045)	
$h=5$	RV				0.720	0.061	0.095	0.122	0.300
					(0.084)	(0.019)	(0.029)	(0.050)	
	RV	0.012			0.720	0.063	0.095	0.122	0.300
		(0.011)			(0.084)	(0.019)	(0.029)	(0.050)	
	RV		-0.045	0.073	0.709	0.103	0.087	0.120	0.303
			(0.023)	(0.020)	(0.082)	(0.029)	(0.028)	(0.050)	
	BV	0.011			0.596	0.074	0.093	0.116	0.355
		(0.011)			(0.064)	(0.018)	(0.024)	(0.041)	
$h=22$	RV				0.525	0.047	0.099	0.138	0.405
					(0.090)	(0.017)	(0.027)	(0.058)	
	RV	0.011			0.526	0.049	0.099	0.138	0.406
		(0.009)			(0.090)	(0.017)	(0.027)	(0.058)	
	RV		-0.028	0.053	0.518	0.076	0.093	0.137	0.408
			(0.021)	(0.019)	(0.088)	(0.026)	(0.026)	(0.058)	
	BV	0.013			0.448	0.052	0.089	0.126	0.420
		(0.009)			(0.072)	(0.015)	(0.023)	(0.047)	
$h=66$	RV				0.329	0.038	0.074	0.175	0.363
					(0.094)	(0.011)	(0.018)	(0.059)	
	RV	0.012			0.329	0.040	0.074	0.174	0.363
		(0.006)			(0.094)	(0.012)	(0.018)	(0.059)	
	RV		-0.023	0.048	0.322	0.064	0.069	0.173	0.366
			(0.015)	(0.015)	(0.094)	(0.019)	(0.018)	(0.059)	
	BV	0.013			0.293	0.040	0.065	0.141	0.349
		(0.007)			(0.087)	(0.011)	(0.017)	(0.049)	

Note: This table uses the same equations as those outlined in Tables 3.4a-3.4c with the addition of the MOVE index. We scale to move index by dividing it by 10^7 in order to have it on the same order of magnitude as RV. Newey-West standard errors are in parentheses with bandwidth selections of 5 ($h=1$), 8 ($h=5$), 42 ($h=22$), and 130 ($h=66$).

Table 3.7b: HAR Estimation with Signed Jumps and Implied Volatility for 10-Year

	$Measure$	β_J	β_J^+	β_J^-	β_{IV}	β_B	β_w	β_m	R^2
$h=1$	RV				2.162	0.079	0.080	-0.044	0.120
					(0.196)	(0.026)	(0.037)	(0.042)	
	RV	0.007			2.163	0.080	0.079	-0.045	0.120
		(0.018)			(0.196)	(0.026)	(0.037)	(0.042)	
	RV		-0.044	0.095	2.137	0.126	0.071	-0.047	0.121
			(0.021)	(0.032)	(0.191)	(0.036)	(0.038)	(0.042)	
	BV	0.008			1.747	0.102	0.091	-0.019	0.186
		(0.017)			(0.153)	(0.026)	(0.032)	(0.036)	
$h=5$	RV				1.802	0.066	0.062	0.063	0.325
					(0.173)	(0.018)	(0.025)	(0.041)	
	RV	0.005			1.803	0.067	0.062	0.063	0.325
		(0.012)			(0.172)	(0.018)	(0.025)	(0.042)	
	RV		-0.042	0.089	1.778	0.110	0.054	0.061	0.328
			(0.014)	(0.023)	(0.168)	(0.023)	(0.024)	(0.041)	
	BV	0.003			1.468	0.077	0.072	0.069	0.391
		(0.012)			(0.143)	(0.017)	(0.025)	(0.037)	
$h=22$	RV				1.469	0.041	0.062	0.093	0.470
					(0.216)	(0.012)	(0.025)	(0.052)	
	RV	0.004			1.470	0.041	0.062	0.092	0.470
		(0.011)			(0.216)	(0.012)	(0.025)	(0.052)	
	RV		-0.030	0.064	1.452	0.072	0.056	0.091	0.474
			(0.008)	(0.019)	(0.210)	(0.017)	(0.024)	(0.052)	
	BV	0.007			1.217	0.048	0.064	0.084	0.468
		(0.011)			(0.174)	(0.012)	(0.023)	(0.049)	
$h=66$	RV				1.084	0.032	0.050	0.121	0.440
					(0.199)	(0.012)	(0.018)	(0.044)	
	RV	0.011			1.084	0.034	0.049	0.120	0.441
		(0.008)			(0.198)	(0.012)	(0.017)	(0.044)	
	RV		-0.019	0.065	1.061	0.064	0.044	0.119	0.444
			(0.007)	(0.019)	(0.196)	(0.018)	(0.016)	(0.044)	
	BV	0.012			0.937	0.035	0.045	0.082	0.396
		(0.009)			(0.186)	(0.011)	(0.016)	(0.040)	

Note: This table uses the same equations as those outlined in Tables 3.4a-3.4c with the addition of the MOVE index. We scale to move index by dividing it by 10^7 in order to have it on the same order of magnitude as RV. Newey-West standard errors are in parentheses with bandwidth selections of 5 ($h=1$), 8 ($h=5$), 42 ($h=22$), and 130 ($h=66$).

Table 3.7c: HAR Estimation with Signed Jumps and Implied Volatility for 30-Year

	$Measure$	β_J	β_J^+	β_J^-	β_{IV}	β_B	β_w	β_m	R^2
$h = 1$	RV				3.327	0.106	0.117	0.134	0.126
					(0.339)	(0.032)	(0.034)	(0.061)	
	RV	-0.017			3.306	0.109	0.118	0.134	0.126
		(0.034)			(0.331)	(0.028)	(0.034)	(0.061)	
	RV		-0.123	0.115	3.162	0.195	0.104	0.127	0.128
			(0.037)	(0.031)	(0.308)	(0.035)	(0.031)	(0.059)	
	BV	-0.014			2.875	0.125	0.121	0.122	0.210
		(0.032)			(0.301)	(0.029)	(0.037)	(0.043)	
$h = 5$	RV				2.865	0.083	0.091	0.216	0.344
					(0.352)	(0.024)	(0.029)	(0.069)	
	RV	-0.009			2.853	0.085	0.091	0.216	0.344
		(0.022)			(0.344)	(0.022)	(0.029)	(0.069)	
	RV		-0.106	0.110	2.722	0.163	0.078	0.210	0.351
			(0.030)	(0.032)	(0.320)	(0.030)	(0.027)	(0.066)	
	BV	-0.011			2.509	0.093	0.095	0.196	0.436
		(0.024)			(0.316)	(0.023)	(0.031)	(0.051)	
$h = 22$	RV				2.424	0.055	0.085	0.239	0.519
					(0.587)	(0.017)	(0.033)	(0.071)	
	RV	-0.011			2.410	0.057	0.086	0.239	0.520
		(0.018)			(0.567)	(0.015)	(0.034)	(0.071)	
	RV		-0.086	0.081	2.309	0.118	0.076	0.234	0.527
			(0.027)	(0.021)	(0.521)	(0.024)	(0.033)	(0.070)	
	BV	-0.009			2.216	0.059	0.083	0.207	0.540
		(0.018)			(0.522)	(0.015)	(0.032)	(0.064)	
$h = 66$	RV				2.061	0.042	0.067	0.224	0.518
					(0.552)	(0.012)	(0.019)	(0.044)	
	RV	-0.003			2.057	0.042	0.067	0.224	0.518
		(0.014)			(0.545)	(0.011)	(0.019)	(0.044)	
	RV		-0.065	0.075	1.966	0.094	0.060	0.221	0.524
			(0.020)	(0.017)	(0.512)	(0.019)	(0.018)	(0.043)	
	BV	-0.001			1.908	0.042	0.062	0.184	0.491
		(0.013)			(0.544)	(0.011)	(0.018)	(0.038)	

Note: This table uses the same equations as those outlined in Tables 3.4a-3.4c with the addition of the MOVE index. We scale to move index by dividing it by 10^7 in order to have it on the same order of magnitude as RV. Newey-West standard errors are in parentheses with bandwidth selections of 5 ($h = 1$), 8 ($h = 5$), 42 ($h = 22$), and 130 ($h = 66$).

Table 3.8a: DM Statistic: HAR-RV vs HAR-RS QLIKE

	5-year	10-year	30-year
$h = 1$	1.372	2.257	0.861
$h = 5$	1.662	3.866	2.254
$h = 22$	1.438	4.197	2.180
$h = 66$	2.014	4.763	3.233

Table 3.8b: DM Statistic: HAR-BV vs HAR-$\Delta J^2\pm$ QLIKE

	5-year	10-year	30-year
$h = 1$	2.883	3.540	2.253
$h = 5$	3.107	3.343	1.295
$h = 22$	3.348	4.548	0.314
$h = 66$	4.750	5.393	2.372

Table 3.8c: DM Statistic: HAR-RV vs HAR-$\Delta J^2\pm$ QLIKE

	5-year	10-year	30-year
$h = 1$	4.429	4.141	2.567
$h = 5$	5.058	3.956	1.927
$h = 22$	6.405	5.048	2.124
$h = 66$	8.102	6.127	3.937

Note: These tables present test statistics under the null hypothesis that the forecasts have equal predictive power. We use Newey-West standard errors when calculating the test statistic to control for autocorrelation in the forecasting errors. The DM test statistic has a standard normal limiting distribution, with a positive test statistic indicating the second model listed outperforms the first model listed.

Table 3.9a: DM Statistic: HAR-RV vs HAR-RS MSE

	5-year	10-year	30-year
$h = 1$	0.869	1.423	0.606
$h = 5$	1.014	2.864	1.535
$h = 22$	0.804	3.425	1.651
$h = 66$	1.673	4.282	2.724

Table 3.9b: DM Statistic: HAR-BV vs HAR-$\Delta J^2\pm$ MSE

	5-year	10-year	30-year
$h = 1$	2.083	1.788	0.420
$h = 5$	2.373	2.102	-0.029
$h = 22$	2.275	3.337	-0.256
$h = 66$	4.022	4.559	1.153

Table 3.9c: DM Statistic: HAR-RV vs HAR-$\Delta J^2\pm$ MSE

	5-year	10-year	30-year
$h = 1$	3.213	2.159	0.796
$h = 5$	3.852	2.388	0.438
$h = 22$	4.811	3.534	0.747
$h = 66$	6.941	5.021	2.334

Note: These tables present test statistics under the null hypothesis that the forecasts have equal predictive power. We use Newey-West standard errors when calculating the test statistic to control for autocorrelation in the forecasting errors. The DM test statistic has a standard normal limiting distribution, with a positive test statistic indicating the second model listed outperforms the first model listed.

Bibliography

An, Y., Y. Hu, and M. Shum (2010): "Estimating first-price auctions with an unknown number of bidders: A misclassification approach," *Journal of Econometrics*, 157(2), 328–341.

Andersen, T. G., T. Bollerslev, and F. X. Diebold (2007): "Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility," *The Review of Economics and Statistics*, 89(4), 701–720.

Asker, J. (2010): "A study of the internal organization of a bidding cartel," *The American Economic Review*, 100(3), 724–762.

Athey, S., and P. A. Haile (2007): "Nonparametric approaches to auctions," *Handbook of Econometrics*, 6, 3847–3965.

Bajari, P., S. Houghton, and S. Tadelis (2006): "Bidding for incomplete contracts: An empirical analysis," Discussion paper, National Bureau of Economic Research.

Bajari, P., and L. Ye (2003): "Deciding between competition and collusion," *Review of Economics and Statistics*, 85(4), 971–989.

Balat, J. (2012): "Highway procurement and the stimulus package: Identification and estimation of dynamic auctions with unobserved heterogeneity," Discussion paper.

Balat, J., P. A. Haile, H. Hong, and M. Shum (2015): "Nonparametric tests for common values at first-price sealed-bid auctions," Discussion paper.

Barndorff-Nielsen, O., S. Kinnebrock, and N. Shephard (2010): "Volatility and Time Series Econometrics: "Essays in Honor of Robert F. Engle, Chapter Measuring Downside Risk-Realised Semivariance"," .

Barndorff-Nielsen, O. E., and N. Shephard (2004): "Power and bipower variation with stochastic volatility and jumps," *Journal of Financial Econometrics*, 2(1), 1–37.

——— (2006): "Econometrics of testing for jumps in financial economics using bipower variation," *Journal of Financial Econometrics*, 4(1), 1–30.

Bayraktar, M., Q. Cui, M. Hastak, and I. Minkarah (2004): "State-of-practice of warranty contracting in the United States," *Journal of Infrastructure Systems*, 10(2), 60–68.

Beenstock, M., and G. Rahav (2002): "Testing Gateway Theory: do cigarette prices affect illicit drug use?," *Journal of Health Economics*, 21(4), 679–698.

Biemer, P., and C. Wiesen (2002): "Measurement error evaluation of self-reported drug use: a latent class analysis of the US National Household Survey on Drug Abuse," *Journal of the Royal Statistical Society: Series A*, 165(1), 97–119.

Biemer, P., and M. Witt (1996): "Estimation of measurement bias in self-reports of drug use with applications to the national household survey on drug abuse," *Journal of Official Statistics*, 12(3), 275–300.

Bray, J., G. Zarkin, C. Ringwalt, and J. Qi (2000): "The relationship between marijuana initiation and dropping out of high school," *Health Economics*, 9(1), 9–18.

Campo, S., I. Perrigne, and Q. Vuong (2003): "Asymmetry in first-price auctions with affiliated private values," *Journal of Applied Econometrics*, 18(2), 179–207.

Chen, X., H. Hong, and D. Nekipelov (2011): "Nonlinear Models of Measurement Errors," *Journal of Economic Literature*, 49(4), 901–37.

Construction, M. D. O. O., and I. Contracting (2006): "Innovative Contracting In Minnesota 2000 To 2005," Discussion paper.

Cooper, R., and T. Ross (1985): "Product Warranties and Double Moral Hazard," *RAND Journal of Economics*, 16(1), 103–113.

Corsi, F. (2009): "A Simple Approximate Long-Memory Model of Realized Volatility," *Journal of Financial Econometrics*, 7(2), 174–196.

Cui, Q., P. Johnson, D. Sharma, and M. Bayraktar (2010): "Determinants of Industry Acceptance for Highway Warranty Contracts: Alabama Case Study," *Journal of Infrastructure Systems*, 16(1), 93–101.

DeSimone, J. (2002): "Illegal drug use and employment," *Journal of Labor Economics*, 20(4), 952–977.

Diebold, F. X. (2015): "Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests," *Journal of Business & Economic Statistics*, 33(1), 1–1.

Diebold, F. X., and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13(3), 253–263.

Figlewski, S., and X. Wang (2000): "Is the 'Leverage Effect' a Leverage Effect?," *Available at SSRN 256109*.

French, M., M. Roebuck, and P. Alexandre (2001): "Illicit Drug Use, Employment, and Labor Force Participation," *Southern Economic Journal*, 68(2), 349–368.

French, M., G. Zarkin, T. Mroz, and J. Bray (1998): "The relationship between drug use and labor supply for young men," *Labour Economics*, 5(4), 385–409.

Genesove, D., and W. P. Mullin (1998): "Testing static oligopoly models: conduct and cost in the sugar industry, 1890-1914," *RAND Journal of Economics*, 29(2), 355–377.

Gharaibeh, N., and H. Shirazi (2009): "Risk-Based Model for Pricing Highway Infrastructure Warranties," *Journal of Infrastructure Systems*, 15(4), 378–382.

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993): "On the relation between the expected value and the volatility of the nominal excess return on stocks," *The Journal of Finance*, 48(5), 1779–1801.

Guerre, E., I. Perrigne, and Q. Vuong (2000): "Optimal Nonparametric Estimation of First-price Auctions," *Econometrica*, 68(3), 525–574.

Hausman, J., W. Newey, H. Ichimura, and J. Powell (1991): "Identification and estimation of polynomial errors-in-variables models," *Journal of Econometrics*, 50(3), 273–295.

Hu, Y. (2008): "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution," *Journal of Econometrics*, 144(1), 27–61.

Hu, Y., and S. Schennach (2008): "Instrumental variable treatment of nonclassical measurement error models," *Econometrica*, 76(1), 195–216.

Hui, S., and S. Walter (1980): "Estimating the error rates of diagnostic tests," *Biometrics*, 36(1), 167–171.

Jofre-Bonet, M., and M. Pesendorfer (2003): "Estimation of a Dynamic Auction Game," *Econometrica*, 71(5), 1443–1489.

Kaestner, R. (1994): "The Effect of Illicit Drug Use on the Labor Supply of Young Adults," *Journal of Human Resources*, 29(1), 126–155.

Kreider, B., J. V. Pepper, et al. (2011): "Identification of Expected Outcomes in a Data Error Mixing Model With Multiplicative Mean Independence," *Journal of Business & Economic Statistics*, 29(1), 49–60.

Kunsch, H. R., et al. (1989): "The Jackknife and the Bootstrap for General Stationary Observations," *The Annals of Statistics*, 17(3), 1217–1241.

Laffont, J.-J., H. Ossard, Q. Vuong, et al. (1995): "Econometrics of First-Price Auctions," *Econometrica*, 63(4), 953–80.

Lewis, G. (2011): "Asymmetric Information, Adverse Selection and Online Disclosure: The Case of eBay Motors," *American Economic Review*, 101(4), 1535–1546.

Lewis, G., and P. Bajari (2011a): "Incentives and Adaptation: Evidence from Highway Procurement in Minnesota," Discussion paper, National Bureau of Economic Research.

——— (2011b): "Procurement contracting with time incentives: Theory and evidence," *The Quarterly Journal of Economics*, 126(3), 1173–1211.

Li, T., I. Perrigne, and Q. Vuong (2000): "Conditionally independent private information in OCS wildcat auctions," *Journal of Econometrics*, 98(1), 129–161.

——— (2002): "Structural Estimation of the Affliated Private Value Auction Model," *RAND Journal of Economics*, 33(2), 171–193.

Lu, J., and I. Perrigne (2008): "Estimating risk aversion from ascending and sealed-bid auctions: the case of timber auction data," *Journal of Applied Econometrics*, 23(7), 871–896.

Lutz, N. A., and V. Padmanabhan (1998): "Warranties, extended warranties, and product quality," *International Journal of Industrial Organization*, 16(4), 463–493.

MacDonald, Z., and S. Pudney (2000): "Illicit drug use, unemployment, and occupational attainment," *Journal of Health Economics*, 19(6), 1089–1115.

Mahajan, A. (2006): "Identification and estimation of regression models with misclassification," *Econometrica*, 74(3), 631–665.

Management, O. D. D. O. C. (2007): "2006 Status of the Warranty Program," Discussion paper.

Mensch, B., and D. Kandel (1988): "Underreporting of substance use in a national longitudinal youth cohort," *Public Opinion Quarterly*, 52(1), 100–124.

Molinari, F. (2010): "Missing treatments," *Journal of Business & Economic Statistics*, 28(1), 82–95.

Norton, E., R. Lindrooth, and S. Ennett (1998): "Controlling for the endogeneity of peer substance use on adolescent alcohol and tobacco use," *Health Economics*, 7(5), 439–453.

Paarsch, H. J., and H. Hong (2006): "An introduction to the structural econometrics of auction data," *MIT Press Books*, 1.

Patton, A. J. (2011): "Volatility forecast comparison using imperfect volatility proxies," *Journal of Econometrics*, 160(1), 246–256.

Patton, A. J., and K. Sheppard (2015): "Good volatility, bad volatility:: Signed jumps and the persistence of volatility," *Review of Economics and Statistics*, 97(3), 683–697.

Polen, M. R., S. Sidney, I. S. Tekawa, M. Sadler, and G. D. Friedman (1993): "Health care use by frequent marijuana smokers who do not smoke tobacco.," *Western Journal of Medicine*, 158(6), 596.

Porter, R., and J. Zona (1993): "Detection of Bid Rigging in Procurement Auctions," *Journal of Political Economy*, 101(3), 518–38.

Roberts, J. (2011): "Can warranties substitute for reputations?," *American Economic Journal: Microeconomics*, 3(3), 69–85.

Sasaki, Y. (2011): "Heterogeneity and Selection in Dynamic Panel Data," Discussion paper.

Schennach, S. (2004): "Estimation of nonlinear models with measurement error," *Econometrica*, 72(1), 33–75.

——— (2007): "Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models," *Econometrica*, 75(1), 201–239.

Silverman, B. W. (1986): *Density estimation for statistics and data analysis*, vol. 26. CRC Press.

Skinner Jr, R. E. (2008): "Highway Design and Construction: The Innovation Challenge," *The Bridge*, 38(2).

Spence, M. (1977): "Consumer misperceptions, product failure and producer liability," *The Review of Economic Studies*, 44(3), 561–572.

Tauchen, G., and H. Zhou (2011): "Realized jumps on financial markets and predicting credit spreads," *Journal of Econometrics*, 160(1), 102–118.

Thompson, B., S. Anderson, J. Russell, and A. Hanna (2002): "Guidelines for warranty contracting for highway construction," *Journal of Management in Engineering*, 18(3), 129–137.

Tiererova, L. (2013): "A Dynamic Model of Bidder Learning in Procurement Auctions," Discussion paper.

Volkow, N. D., R. D. Baler, W. M. Compton, and S. R. Weiss (2014): "Adverse health effects of marijuana use," *New England Journal of Medicine*, 370(23), 2219–2227.

Appendix A

Appendix for Chapter 1

A.1 Bid Function for Procurement Auction with Warranty

We consider the expected profit function of player i:

$$E[\pi_i|c_{1i}] = (b_i - c_{1i}(1+e_i) - ProbCA(e_i)\delta\mu_c)Pr(b_i < b_{-i})$$

where $-i$ represents all other bidders. As a reminder, c_{2i} is a second draw made in the future if one needs to do corrective action, and since players are bidding based on expected costs it will be treated as a constant, μ_c, under the further assumption that both c_{1i} and c_{2i} are independently drawn from the same distribution. e_i is the extra effort of the project. I will assume that the extra effort choice is binary, $e_i \in \{0, \bar{e}\}$.

Thus a player's expected costs when she chooses $e_i = 0$ are $c_{1i}+ProbCA(e_i = 0)\delta\mu_c$ and a player's costs when choosing high effort are $c_{1i}(1+\bar{e})+ProbCA(e_i = \bar{e})\delta\mu_c$. The player can solve the maximization with respect to e_i while abstracting away from the choice of b_i. A player will choose their optimal effort $e_i^* \in \{0, \bar{e}\}$ following the rule:

$$e_i^* = \begin{cases} 0 & \text{if } c_{1i}(1+\bar{e}) - c_{1i} > ProbCA(e_i = 0)\delta\mu_c \\ \bar{e} & \text{otherwise} \end{cases}$$

I will refer to the following object as the cutoff point where players are indifferent between the effort choices:

$$c^* \equiv \frac{(ProbCA(e_i = 0)\delta\mu_c}{\bar{e}}. \tag{A.1}$$

Assuming other bidders use an invertible bid function in equilibrium, I can rewrite the expected profit function as:

$$E[\pi_i|c_{1i}] = (b_i - c_{1i}(1+e_i) - ProbCA(e_i)\delta\mu_c)Pr(b_i < \beta(c_{-1i}))$$

$$= (b_i - c_{1i}(1+e_i) - ProbCA(e_i)\delta\mu_c)[1 - F(\beta^{-1}(b_i))]^{n-1}$$

Now taking the first order conditions with respect to b_i yields:

$$\frac{\partial E[\pi_i|c_{1i}]}{\partial b_i} = [1-F(\beta^{-1}(b_i))]^{n-1} + (b_i - c_{1i}(1+e_i) - ProbCA(e_i)\delta\mu_c)\frac{\partial[1 - F(\beta^{-1}(b_i))]^{n-1}}{\partial b_i} = 0$$

Assuming a symmetric equilibrium, I can substitute c_{1i} for $\beta^{-1}(b_i)$ and $\beta(c_{1i})$ for b_i. Additionally, I know $\frac{\partial[1-F(\beta^{-1}(b_i))]^{n-1}}{\partial b_i} = \frac{(n-1)[1-F(\beta^{-1}(b_i))]^{n-2}(-f(\beta^{-1}(b_i)))}{\beta'(\beta^{-1}(b_i)}$. Making these substitutions gives me:

$$[1-F(c_{1i})]^{n-1} + (\beta(c_{1i}) - c_{1i}(1+e_i) - ProbCA(e_i)\delta\mu_c)(N-1)F(c_{1i}))^{n-2}(-f(c_{1i})))\frac{1}{\beta'(c_{1i})} = 0$$

Rearranging terms:

$$[1 - F(c_{1i})]^{n-1}\beta'(c_{1i}) + \beta(c_{1i})(n-1)[1 - F(c_{1i})]^{n-2}(-f(c_{1i})) =$$

$$(c_{1i}(1+e_i) + ProbCA(e_i)\delta\mu_c)(n-1)[1 - F(c_{1i})]^{n-2}(-f(c_{1i})) \tag{A.2}$$

Which is the same as:

$$\frac{d[1-F(c_{1i})]^{n-1}\beta(c_{1i})}{dc_1} = (c_{1i}(1+e_i) + ProbCA(e_i)\delta\mu_c)(n-1)[1-F(c_{1i})]^{n-2}(-f(c_{1i}))$$

Integrating both sides (with the RHS using integration by parts) and imposing the boundary condition that $\beta(\bar{c}) = \bar{c} + ProbCA(e_i = 0)\delta\mu_c$, because I assume a nondegenerate case such that there exists some part of the distribution of c_{1i} such that $e_i = 0$ is realized:

$$[1-F(c_{1i})]^{n-1}\beta(c_{1i}) = (c_{1i}(1+e_i) + ProbCA(e_i)\delta\mu_c)[1-F(c_{1i})]^{n-1} + \int_{c_{1i}}^{\bar{c}} (1+e_i)[1-F(x)]^{n-1}dx \quad \text{(A.3)}$$

Dividing both sides by $[1-F(c_{1i})]^{n-1}$ gives me the familiar equation that shows the bid as an additive function of the signal and some markdown term:

$$\beta(c_{1i}) = c_{1i}(1+e_i) + ProbCA(e_i)\delta\mu_c + \frac{(1+e_i) \times \int_{c_{1i}}^{\bar{c}}[1-F(x)]^{N-1}dx}{[1-F(c_{1i})]^{n-1}} \quad \text{(A.4)}$$

This function applies to all costs such that $c_{1i} > c^*$, which means I am also able to make the following substitution in terms of effort selection:

$$\beta(c_{1i}) = c_{1i} + ProbCA(e_i = 0)\delta\mu_c + \frac{\int_{c_{1i}}^{\bar{c}}[1-F(x)]^{n-1}dx}{[1-F(c_{1i})]^{n-1}} \qquad \text{for} \quad c_{1i} > c^* \quad \text{(A.5)}$$

Next, consider equation A.2 where we know the solution at $\beta(c^*)$ is continuous, due to the continuity of total expected costs. Recall that if $c_{1i} < c^*$, bidders choose $e_i = \bar{e}$. Thus, the solution gives:

$$\beta(c_{1i}) = c_{1i}(1+e_i) + ProbCA(e_i = \bar{e})\delta\mu_c + \frac{(1+\bar{e}) \times \int_{c_{1i}}^{c^*}[1-F(x)]^{n-1}dx}{[1-F(c_{1i})]^{n-1}} \quad \text{(A.6)}$$

$$+\frac{\int_{c^*}^{\bar{c}}[1-F(x)]^{n-1}dx}{[1-F(c_{1i})]^{n-1}} \qquad \text{for} \quad c_{1i} < c^*$$

Furthermore, if we impose the assumption put forth in the paper that high effort work is related to the case where contractors guarantee they do not have to undertake higher effort work ($ProbCA(e_i = \bar{e}) = 0$), the equation simplifies to:

$$\beta(c_{1i}) = c_{1i}(1+e_i) + \frac{(1+\bar{e}) \times \int_{c_{1i}}^{c^*}[1-F(x)]^{n-1}dx}{[1-F(c_{1i})]^{n-1}} \tag{A.7}$$
$$+\frac{\int_{c^*}^{\bar{c}}[1-F(x)]^{n-1}dx}{[1-F(c_{1i})]^{n-1}} \qquad \text{for} \quad c_{1i} < c^*$$

Taken together, the full equilibrium bid function is defined as:

$$\beta(c_{1i}) = c_{1i} + ProbCA(e_i = 0)\delta c_{2i} + \frac{\int_{c_{1i}}^{\bar{c}}[1-F(x)]^{n-1}dx}{[1-F(c_{1i})]^{n-1}} \qquad \text{for} \quad c_{1i} > c^* \tag{A.8}$$
$$\beta(c_{1i}) = c_{1i}(1+\bar{e}) + \frac{(1+\bar{e}) \times \int_{c_{1i}}^{c^*}[1-F(x)]^{n-1}dx}{[1-F(c_{1i})]^{n-1}}$$
$$+\frac{\int_{c^*}^{\bar{c}}[1-F(x)]^{n-1}dx}{[1-F(c_{1i})]^{n-1}} \qquad \text{for} \quad c_{1i} < c^*$$

A.2 Identification of Effort and Probability of Corrective Action

Next, let me define the percentage of all warranty auctions needing corrective action as:

$$\overline{ProbCA}_n = ProbCA(e_i = 0) * (1 - H_n(c^*)) \tag{A.9}$$

where $H_n()$ represents the the distribution of winning bids, since I only observe the winners needing to undertake corrective action. Due to the monotonicity of the bid function, this is an ordered statistic and the same as the distribution of the lowest cost. I use subscript n to emphasize the dependence of this ordered statistic on the total number of bidders. I specifically look at $1-H_n(c^*)$ because that is the fraction of winning bidders who choose to exert no extra effort.

Under the assumption that I observe auctions where the number of bidders are not the same. I call two arbitrary number of bidders in different auctions n_1 and n_2 for convenience. This allows me to identify the remaining terms from the new data. To see this, I consider the following two equations:

$$\overline{ProbCA}_{n_1} = ProbCA(e_i = 0) * (1 - H_{n_1}(c^*)) \tag{A.10}$$

and

$$\overline{ProbCA}_{n_2} = ProbCA(e_i = 0) * (1 - H_{n_2}(c^*)) \tag{A.11}$$

Recall, $H_{n_1}(c^*)$ represents the probability that the winner's cost draw is less than c^* in the case that $n = n_1$. This is the same as 1 minus the probability that the winner's cost draw is greater than c^* in the case that $n = n_1$. With i.i.d. draws and monotonic bid functions, as I have in my setting, the probability that the winner's cost draw is greater than c^* is the same as saying the probability that all bidders' cost draws are greater than c^*, or $[1 - F(c^*)]^{n_1}$. making these substitutions:

$$\overline{ProbCA}_{n_1} = ProbCA(e_i = 0) * (1 - [1 - [1 - F(c^*)]^{n_1}])$$

which simplifies to

$$\overline{ProbCA}_{n_1} = ProbCA(e_i = 0) * [1 - F(c^*)]^{n_1}$$

Similarly,

$$\overline{ProbCA}_{n_2} = ProbCA(e_i = 0) * [1 - F(c^*)]^{n_2} \tag{A.12}$$

Now let's divide the previous two equations to arrive at:

$$\frac{\overline{ProbCA}_{n_1}}{\overline{ProbCA}_{n_2}} = \frac{[1 - F(c^*)]^{n_1}}{[1 - F(c^*)]^{n_2}}$$

This in turn simplifies to

$$\frac{\overline{ProbCA}_{n_1}}{\overline{ProbCA}_{n_2}} = [1 - F(c^*)]^{n_1 - n_2}$$

The left hand side of this equation is just some constant that comes from the data, and the RHS is monotonic in c^*, thus uniquely identifying that value. With the proper c^* pinned down, any of the corrective action equations, say equation A.10, determine $ProbCA(e_i = 0)$ since all other elements in the equation are known. Similarly, with $ProbCA(e_i = 0)$ determined, $\bar{e_i}$ is uniquely determined given that c^* and μ_c are known from:

$$c^* \equiv \frac{(ProbCA(e_i = 0)\delta\mu_c}{\bar{e}}. \tag{A.13}$$

A.3 Monte Carlo Simulation

I include this brief simulation exercise to demonstrate the properties of my estimator. I consider auctions with $n = 2, 3$, and 4. This simulation assumes

$L = 200$ and goes through 1000 trials. Following the simulation exercise in Li, Perrigne, and Vuong (2002), I draw private signals from the triangular distribution where:

$$f_c() = \begin{cases} 4c & \text{if } 0 \leq c \leq .5 \\ 4 - 4c & \text{if } .5 < c \leq 1 \end{cases}$$

Unlike Li, Perrigne, and Vuong (2002), observed auction heterogeneity is an important aspect of my estimation and I need to alter these draws in such a way that I can test the performance of my homogenization procedure. To do so, I generate two covariates as a simple example. The homogonization regression, because it is done by OLS, has well known properties that cause me the least concern for the entire estimation procedure. Table 1.6 summarizes the results. Figures 1.3a and 1.3b plot nonparametric CDF and pdf estimates that are performed directly on the (unobservable to the econometrician) cost draws. I do this to better understand where any estimation error may be coming from. Next, I complete the full analysis where I use the inverse bid function to back out pseudocosts. Figures 1.4a and 1.4b plot the results. Lastly, I use 3 moments related to corrective action (one for each n) to estimate the remaining parameters. My results are displayed in Table 1.7.

A.4 Discussion of δ

My data include a measure of the estimated costs of corrective action, as reported to MDOT by the contractor. Before using these data directly, I was told by engineers at MDOT that this was a very noisy measure and they did not believe some of the reported amounts (some of which included \$1). Not wanting

to throw away data that the engineers considered plausible, I used their industry expertise by following suggestions of how to process the data in order to get the most reliable self-reported value. Upon their suggestion, I first ignored all reports less than $1000, as the tasks that required corrective action were always more costly than this.[1] I will treat these as random error events.

Next, I divided the remaining reported estimates by the calculated average expected costs (accounting for covariates) for each auction needing corrective action. Note that I am able to do this due to the assumption that c_{1i} and c_{2i} are independently drawn. Lastly, I take the median of all ratios in order to further prevent unreasonable outliers from overly affecting my estimate. This procedure leads to a result which matches the basic estimates put forth by the engineers at MDOT, $\delta = .08$. Similar to how other papers have relied on industry experts and insight from from testimony (see Genesove and Mullin (1998)) or general market institutions (see Laffont, Ossard, Vuong, et al. (1995)) in order to pin down certain parameters, I consider my calculation from self-reported estimates and conversations with the engineers at MDOT to be the best I can do given my data limitations.

[1] One explanation for why there were so many worthless entries was that the internal computer system used to track projects forced engineers to enter a figure before being able to close out a project. This led to engineers calling the construction firms at inopportune times and sometimes getting unreasonable responses from people at the contracting firm who were unwilling to take the time to accurately account for internal costs. Additionally, it was mentioned that some firms may not want to reveal any more information about their costs to the state.

Appendix B

Appendix for 2

B.1 Proofs of Theorems

In this appendix we give brief proofs of our two theorems in the model section.

B.1.1 Proof of Theorem 1.1

The first main equation we have is,

$$\sum_{H_t} \omega\left(H_t\right) \Pr\left(D_t, H_t, D_{t-1}, H_{t-1}, D_{t-2}\right)$$

$$= \sum_{D_t^*} \Pr\left(D_t | D_t^*, D_{t-1}, H_{t-1}, D_{t-2}\right) \times E\left[\omega\left(H_t\right) | D_t^*, D_{t-1}, H_{t-1}, D_{t-2}\right] \times \Pr\left(D_t^*, D_{t-1}, H_{t-1}, D_{t-2}\right)$$

$$= \sum_{D_t^*} \Pr\left(D_t | D_t^*, D_{t-1}\right) \times E\left[\omega\left(H_t\right) | D_t^*, D_{t-1}, H_{t-1}\right] \times \Pr\left(D_t^*, D_{t-1}, H_{t-1}, D_{t-2}\right)$$

according to Assumption 1.1 and 1.2. Similarly,

$$\Pr(D_t, D_{t-1}, H_{t-1}, D_{t-2})$$

$$= \sum_{D_t^*} \Pr\left(D_t | D_t^*, D_{t-1}, H_{t-1}, D_{t-2}\right) \times \Pr\left(D_t^*, D_{t-1}, H_{t-1}, D_{t-2}\right)$$

$$= \sum_{D_t^*} \Pr\left(D_t | D_t^*, D_{t-1}\right) \times \Pr\left(D_t^*, D_{t-1}, H_{t-1}, D_{t-2}\right)$$

Then we put the probabilities on both sides into matrix forms. For the first equation, the left-hand side probabilities could be written as,

$$L_{\omega(H_t)(D_t,H_t,d_{t-1},h_{t-1},D_{t-2})}$$

$$= \begin{bmatrix} g_5(1,1) & g_5(1,0) \\ g_5(0,1) & g_5(0,0) \end{bmatrix}$$

where $g_5(i,,j) = \sum_{H_t} \omega(H_t) \Pr(D_t = i, H_t, d_{t-1}, h_{t-1}, D_{t-2} = j)$.

And for the right-hand side probabilities,

$$L_{D_t^*,d_{t-1},h_{t-1},D_{t-2}}$$

$$= \begin{bmatrix} g_6(1,1) & g_6(1,0) \\ g_6(0,1) & g_6(0,0) \end{bmatrix}$$

where, $g_6(i,j) = \Pr(D_t^* = i, d_{t-1}, h_{t-1}, D_{t-2} = j)$.

$$L_{D_t,d_{t-1},h_{t-1},D_{t-2}}$$

$$= \begin{bmatrix} g_7(1,1) & g_7(1,0) \\ g_7(0,1) & g_7(0,0) \end{bmatrix}$$

where, $g_7(i,j) = \Pr(D_t = i, d_{t-1}, h_{t-1}, D_{t-2} = j)$.

And,

$$D_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}}$$

$$= \begin{bmatrix} E[1] & \\ & E[0] \end{bmatrix}$$

where, $E[i] = E_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}}[i] = E[\omega(H_t)|D_t^* = i, d_{t-1}, h_{t-1}]$,

Thus, equation (2.3) is equivalent to

$$L_{\omega(H_t)(D_t,H_t,d_{t-1},h_{t-1},D_{t-2})} = L_{D_t|D_t^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}} \times L_{D_t^*,d_{t-1},h_{t-1},D_{t-2}}$$

And for the second equation,

$$L_{D_t,d_{t-1},h_{t-1},D_{t-2}}$$

$$= \begin{bmatrix} g_8(1,1) & g_8(1,0) \\ g_8(0,1) & g_8(0,0) \end{bmatrix}$$

where,$g_8(i,j) = \Pr(D_t = i|D_t^* = j, d_{t-1})$.

Thus, we are able to rewrite the second main equations as,

$$L_{D_t,d_{t-1},h_{t-1},D_{t-2}} = L_{D_t|D_t^*,d_{t-1}} \times L_{D_t^*,d_{t-1},h_{t-1},D_{t-2}}.$$

Given Assumption 1.3 which ensures invertibility of the left-hand side matrix in equation (2.5), we take the inverse of both sides in this equation and the equation above therefore becomes,

$$L^{-1}_{D_t,d_{t-1},h_{t-1},D_{t-2}} = L^{-1}_{D_t^*,d_{t-1},h_{t-1},D_{t-2}} \times L^{-1}_{D_t|D_t^*,d_{t-1}}.$$

Finally, we right-multiply each side of the equation with the corresponding side in equation(2.4), to get equation (2.6).

$$L_{\omega(H_t)(D_t,H_t,d_{t-1},h_{t-1},D_{t-2})} \times L^{-1}_{D_t,d_{t-1},h_{t-1},D_{t-2}}$$

$$= L_{D_t|D_t^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}} \times L_{D_t^*,d_{t-1},h_{t-1},D_{t-2}} \times$$

$$L^{-1}_{D_t^*,d_{t-1},h_{t-1},D_{t-2}} \times L^{-1}_{D_t|D_t^*,d_{t-1}}$$

$$= L_{D_t|D_t^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}} \times L^{-1}_{D_t|D_t^*,d_{t-1}}$$

In the last line of the equation, the RHS matrcies include the misclassification error which is of our central interest, and they could be recovered via a eigenvector-eigenvalue decomposition of the LHS matrix. But as we know from basic matrix arithmetic, the exact position of each eigenvector in the matrix is not determined. In order to reconcile this problem, we introduce Assumption 1.4 and 1.5 which ensure the most reasonable ordering of eigenvectors from an economic point of view. Once the order of the columns of eigenvectors is determined, the misclassification error is uniquely identified and could be estimated. Therefore, Theorem 1.1 is proved.

B.1.2 Proof of Theorem 2.1

Following the similar logic as in the previous proof, we firstly write down the two main equations for the general approach,

$$\sum_{H_t} \omega\left(H_t\right) \Pr\left(D_{t+1}, D_t, H_t, D_{t-1}, H_{t-1}, D_{t-2}, D_{t-3}\right)$$

$$= \sum_{D_t^*, D_{t-1}^*} \Pr\left(D_{t+1}, D_t | D_t^*, D_{t-1}^*, D_{t-1}, D_{t-2}, D_{t-3}, H_{t-1}\right) \times$$

$$E\left[\omega\left(H_t\right) | D_t^*, D_{t-1}^*, D_{t+1}, D_t, D_{t-1}, H_{t-1}, D_{t-2}, D_{t-3}\right] \times$$

$$\Pr\left(D_t^*, D_{t-1}^*, D_{t-1}, H_{t-1}, D_{t-2}, D_{t-3}\right)$$

$$= \sum_{D_t^*, D_{t-1}^*} \Pr\left(D_{t+1}, D_t | D_t^*, D_{t-1}^*, D_{t-1}\right) \times E\left[\omega\left(H_t\right) | D_t^*, D_{t-1}^*, D_{t-1}, H_{t-1}\right] \times$$

$$\Pr\left(D_t^*, D_{t-1}^*, D_{t-1}, H_{t-1}, D_{t-2}, D_{t-3}\right).$$

and the second equality follows from Assumption 2.1 and 2.2. Similarly, we write down the second main equation as,

$$\Pr(D_{t+1}, D_t, D_{t-1}, H_{t-1}, D_{t-2}, D_{t-3})$$

$$= \sum_{D_t^*, D_{t-1}^*} \Pr(D_{t+1}, D_t | D_t^*, D_{t-1}^*, D_{t-1}, H_{t-1}, D_{t-2}, D_{t-3}) \times \Pr(D_t^*, D_{t-1}^*, D_{t-1}, H_{t-2}, D_{t-2}, D_{t-3})$$

$$= \sum_{D_t^*, D_{t-1}^*} \Pr(D_{t+1}, D_t | D_t^*, D_{t-1}^*, D_{t-1}) \times \Pr(D_t^*, D_{t-1}^*, D_{t-1}, H_{t-2}, D_{t-2}, D_{t-3})$$

Now we put all the probabilities in the equations into matrix forms. For the first one,

$$L_{\omega(H_t)(D_{t+1}, D_t, H_t, d_{t-1}, h_{t-1}, D_{t-2}, D_{t-3})}$$

$$= \begin{bmatrix} g_1(1,1,1,1) & g_1(1,1,1,0) & g_1(1,1,0,1) & g_1(1,1,0,0) \\ g_1(1,0,1,1) & g_1(1,0,1,0) & g_1(1,0,0,1) & g_1(1,0,0,0) \\ g_1(0,1,1,1) & g_1(0,1,1,0) & g_1(0,1,0,1) & g_1(0,1,0,0) \\ g_1(0,0,1,1) & g_1(0,0,1,0) & g_1(0,0,0,1) & g_1(0,0,0,0) \end{bmatrix}$$

where $g_1(i,j,r,k) = \sum_{H_t} \omega(H_t) \Pr(D_{t+1}=i, D_t=j, H_t, D_{t-1}, H_{t-1}, D_{t-2}=r, D_{t-3}=k)$, for any i,j,r,k;

$$L_{D_{t+1},D_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3}}$$

$$= \begin{bmatrix} g_2(1,1,1,1) & g_2(1,1,1,0) & g_2(1,1,0,1) & g_2(1,1,0,0) \\ g_2(1,0,1,1) & g_2(1,0,1,0) & g_2(1,0,0,1) & g_2(1,0,0,0) \\ g_2(0,1,1,1) & g_2(0,1,1,0) & g_2(0,1,0,1) & g_2(0,1,0,0) \\ g_2(0,0,1,1) & g_2(0,0,1,0) & g_2(0,0,0,1) & g_2(0,0,0,0) \end{bmatrix}$$

where $g_2(i,j,r,k) = \Pr(D_{t+1}=i, D_t=j, d_{t-1}, h_{t-1}, D_{t-2}=r, D_{t-3}=k)$,for any i,j,r,k;

$$L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}}$$

$$= \begin{bmatrix} g_3(1,1|1,1) & g_3(1,1|1,0) & g_3(1,1|0,1) & g_3(1,1|0,0) \\ g_3(1,0|1,1) & g_3(1,0|1,0) & g_3(1,0|0,1) & g_3(1,0|0,0) \\ g_3(0,1|1,1) & g_3(0,1|1,0) & g_3(0,1|0,1) & g_3(0,1|0,0) \\ g_3(0,0|1,1) & g_3(0,0|1,0) & g_3(0,0|0,1) & g_3(0,0|0,0) \end{bmatrix}$$

where $g_3(i,j|r,k) = \Pr(D_{t+1}=i, D_t=j|D_t^*=r, D_{t-1}^*=k, d_{t-1})$, for any i,j,r,k;

$$D_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}}$$

$$= \begin{bmatrix} E[1,1] & & & \\ & E[1,0] & & \\ & & E[0,1] & \\ & & & E[0,0] \end{bmatrix}$$

where, $E[i,j] = E_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}}[i,j] = E\left[\omega(H_t)|D_t^*=i, D_{t-1}^*=j, d_{t-1}, h_{t-1}\right]$, for any i,j;

$$L_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}}$$

$$= \begin{bmatrix} g_4(1,1,1,1) & g_4(1,1,1,0) & g_4(1,1,0,1) & g_4(1,1,0,0) \\ g_4(1,0,1,1) & g_4(1,0,1,0) & g_4(1,0,0,1) & g_4(1,0,0,0) \\ g_4(0,1,1,1) & g_4(0,1,1,0) & g_4(0,1,0,1) & g_4(0,1,0,0) \\ g_4(0,0,1,1) & g_4(0,0,1,0) & g_4(0,0,0,1) & g_4(0,0,0,0) \end{bmatrix}$$

where $g_4(i,j,r,k) = \Pr(D_t^* = i, D_{t-1}^* = j, d_{t-1}, h_{t-1}, D_{t-2} = r, D_{t-3} = k)$, for any i, j, r, k. Thus, the matrix notation would be written as:

$$L_{\omega(H_t)(D_{t+1},D_t,H_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3})} = L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}} \times L_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}}$$

and corresponding to a degenerated $\omega(\cdot) = 1$

$$L_{D_{t+1},D_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} = L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}} \times L_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}}$$

Given Assumption 2.3 which ensures invertibility of the left-hand side matrix in equation (2.13), we take the inverse of both sides and obtain the new equation,

$$L^{-1}_{D_{t+1},D_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} = L^{-1}_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} \times L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}}$$

and then we right-multiply each side with the corresponding side in equation(2.12). Finally, we get equation (2.14).

$$L_{\omega(H_t)(D_{t+1},D_t,H_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3})} \times L^{-1}_{D_{t+1},D_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3}}$$

$$= L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}} \times L_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} \times$$

$$L^{-1}_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} \times L^{-1}_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}}$$

$$= L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}} \times L^{-1}_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}}$$

Similarly as in the basic case, this equation above could help us to recover the eigenvectors that represents the misclassification errors of our central interest on the right-hand side. in addition, Assumption 2.4 and 2.5 enable us

to determine the correct ordering of each eigenvector, therefore completes the proof for Theorem 2.1.

B.2 Simulation: Basic Approach

B.2.1 Data Generation

Data of size N is generated according to the following procedure for 1000 repetition:

Step 1 Generate D_{t-1} using the marginal distribution $\Pr(D_{t-1})$ from the true data.

Step 2 Conditional on $D_{t-1} \in \{0, 1\}$ simultaneously generate (D_t^*, D_{t-2}) using the joint distribution of $\Pr(D_t^*, D_{t-2}|D_{t-1})$.

Step 3 Conditional on D_{t-1} and D_t^*, generate D_t using $\Pr(D_t|D_t^*, D_{t-1})$.

Step 4 Generate the health condition data, H_t, using $\Pr(H_t = 1|D_t^*, D_{t-1})$.

In order to get the distribution of H_t, it is better to assume it takes on binary values. This is because only $E[H_t|D_t^*, D_{t-1}, H_{t-1}]$ can be identified using our current method. Therefore a binary H_t could give us accurate estimation of $\Pr(H_t|D_t^*, D_{t-1}, H_{t-1})$. In our real estimation, H_t takes four possible values, $(1, 2.5, 4, 5)$. To get the parameters for simulation, we assume $H_\tau = 0$ if $H_\tau \leq 3$ originally and 1 otherwise.

Thus we have a data set containing information about $(H_t, H_{t-1}, D_t, D_t^*, D_{t-1,}D_{t-2})$. We use our identification method to estimate $\Pr(D_t|D_t^*, D_{t-1})$ and compare the results with the underlying true values.

B.2.2 Parametrization

The underlying parameter values now are given in the following equations and tables:

$$\Pr(D_t^* = 1) = 0.5000, \Pr(D_{t-1} = 1) = 0.5000$$

$$\Pr(D_t^* = 1|D_{t-1}^* = 1) = 0.5000, \Pr(D_t^* = 1|D_{t-1}^* = 0) = 0.5000$$

Table B.1: Joint Distribution of $f(D_t^*, D_{t-2}|D_{t-1})$

$f(D_t^*, D_{t-2}\|D_{t-1})$	$D_{t-1} = 1$	$D_{t-1} = 0$
$D_t^* = 1, D_{t-2} = 1$	0.2000	0.3000
$D_t^* = 1, D_{t-2} = 0$	0.1000	0.2000
$D_t^* = 0, D_{t-2} = 1$	0.1000	0.1000
$D_t^* = 0, D_{t-2} = 0$	0.6000	0.4000

Table B.2: Conditional Distribution of D_t and H_t

	$D_t^* = 1$	$D_t^* = 0$
$Pr(D_t = 1\|D_t^*, D_{t-1} = 1)$	0.2000	0.2500
$Pr(D_t = 1\|D_t^*, D_{t-1} = 0)$	0.0500	0.1000
$E(H_t\|D_t^*, H_{t-1} = 1)$	0.6000	0.5000
$E(H_t\|D_t^*, H_{t-1} = 0)$	0.9000	0.8000

B.3 Simulation: General Approach

B.3.1 Data Generation

Data of size N is generated according to the following procedure for 1000 repetition[1]:

Step 1 Generate D_{t-1} using the marginal distribution $\Pr(D_{t-1})$ from the true data.

Step 2 Conditional on $D_{t-1} \in \{0, 1\}$, we use the joint distribution of $\Pr(D_t^*, D_{t-1}^*, D_{t-2}, D_{t-3}|d_{t-1})$ to generate simultaneously $(D_t^*, D_{t-1}^*, D_{t-2}, D_{t-3})$ given d_{t-1}.

Step 3 We use the estimated conditional probability $\Pr(D_\tau|D_\tau^*, D_{\tau-1}^*, D_{\tau-1})$ to generate D_t .

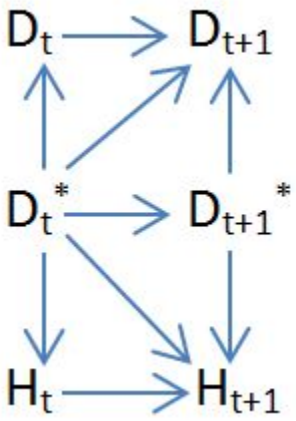
Step 4 Using the conditional distribution of $\Pr(D_{t+1}^*|D_t^*)$ we can generate D_{t+1}^*.

Step 5 Similar to that in step 3, we use the estimated conditional probability $\Pr(D_\tau|D_\tau^*, D_{\tau-1}^*, D_{\tau-1})$ to generate D_{t+1} .

Step 6 The next step is to generate H_τ given $H_{\tau-1}$, D_τ^* and $D_{\tau-1}^*$ for $\tau \in \{t, t+1\}$. This can be achieved by applying the estimated result of $\Pr(H_\tau|D_\tau^*, D_{\tau-1}^*, H_{\tau-1})$.

[1]The simulation is executed in MATLAB, and can be obtained upon request.

Ideally the data should be generated as an evolution from past periods to future ones, as is shown in the following flow chart:



where first a joint distribution of $(H_{t-4}, D_{t-4}, D^*_{t-4})$ is assumed, then D^*_τ is generated from $\Pr(D^*_\tau | D^*_{\tau-1})$, D_τ is generated from $\Pr(D_\tau | D^*_\tau, D^*_{\tau-1}, D_{\tau-1})$, and H_τ is generated from $\Pr(H_\tau | D^*_\tau, D^*_{\tau-1}, H_{\tau-1})$. Yet by using this process, we cannot effectively control the invertibility of the LHS matrix within a limited sample size. Therefore we adopt a more direct approach described above, without violating assumptions we made to get nonsingular matrices.

B.3.2 Parametrization

The underlying parameter values are given in the following equations and in tables:

$$\Pr(D^*_t = 1) = 0.5000, \Pr(D_{t-1} = 1) = 0.5000$$

$$\Pr(D_t^* = 1|D_{t-1}^* = 1) = 0.5000, \Pr(D_t^* = 1|D_{t-1}^* = 0) = 0.5000$$

Table B.3: Joint Distribution of $f(D_t^*, D_{t-1}^*, D_{t-2}, D_{t-3}|D_{t-1} = 1)$

$f(D_t^*, D_{t-1}^*, D_{t-2}, D_{t-3}\|D_{t-1} = 1)$	$D_{t-2} = 1$, $D_{t-3} = 1$	$D_{t-2} = 1$, $D_{t-3} = 0$	$D_{t-2} = 0$, $D_{t-3} = 1$	$D_{t-2} = 0$, $D_{t-3} = 0$
$D_t^* = 1, D_{t-1}^* = 1$	0.1500	0.0500	0.0250	0.0250
$D_t^* = 1, D_{t-1}^* = 0$	0.0250	0.1750	0.0250	0.0250
$D_t^* = 0, D_{t-1}^* = 1$	0.0250	0.0250	0.1500	0.0500
$D_t^* = 0, D_{t-1}^* = 0$	0.0250	0.0250	0.0250	0.1750

Table B.4: Joint Distribution of $f(D_t^*, D_{t-1}^*, D_{t-2}, D_{t-3}|D_{t-1} = 0)$

$f(D_t^*, D_{t-1}^*, D_{t-2}, D_{t-3}\|D_{t-1} = 0)$	$D_{t-2} = 1$, $D_{t-3} = 1$	$D_{t-2} = 1$, $D_{t-3} = 0$	$D_{t-2} = 0$, $D_{t-3} = 1$	$D_{t-2} = 0$, $D_{t-3} = 0$
$D_t^* = 1, D_{t-1}^* = 1$	0.1500	0.0500	0.0250	0.0250
$D_t^* = 1, D_{t-1}^* = 0$	0.0250	0.1750	0.0250	0.0250
$D_t^* = 0, D_{t-1}^* = 1$	0.0250	0.0250	0.1500	0.0500
$D_t^* = 0, D_{t-1}^* = 0$	0.0250	0.0250	0.0250	0.1750

Table B.5: Conditional Distribution of D_t and H_t

	$D_t^* = 1$, $D_{t-1}^* = 1$	$D_t^* = 1$, $D_{t-1}^* = 0$	$D_t^* = 0$, $D_{t-1}^* = 1$	$D_t^* = 0$, $D_{t-1}^* = 0$
$Pr(D_t = 1\|D_t^*, D_{t-1}^*, D_{t-1} = 1)$	0.8000	0.1000	0.7000	0.1000
$Pr(D_t = 1\|D_t^*, D_{t-1}^*, D_{t-1} = 0)$	0.6000	0.1000	0.5000	0.0500
$E(H_t\|D_t^*, D_{t-1}^*, H_{t-1} = 1)$	0.8000	0.5000	0.3000	0.1000
$E(H_t\|D_t^*, D_{t-1}^*, H_{t-1} = 0)$	0.6000	0.4000	0.2000	0.0500

B.4 An Alternative Way of Dealing with H_τ: Combining $H_\tau = 1, 2, 3$

B.4.1 Basic Approach

In this section we combine all the samples with at least "good" health conditions in their surveys answers. Specifically, we redefine $H_\tau = 2.5$ if it originally is $1, 2$

or 3. The results from this second estimation are shown in Table B.6. In addition, we re-estimate everything conditional on gender covariates, and the results are presented in Table B.7 and Table B.8. Compared with those in Section 4, the results here are not very different, except for the estimation for male subgroup. It can be seen that the point estimates for misclassification errors are the same for $D_t^* = 1, D_{t-1} = 1$ case and for $D_t^* = 0, D_{t-1} = 1$ case, which does not make much sense. Nonetheless, the bootstrap means and medians provide more useful information about this subgroup, using the newly defined subsample.

Table B.6: Estimation Results: Basic Approach, Combining $H_\tau = 1, 2$ and 3

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 1)$	0.7992	0.8475	0.8394	0.1152
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 1)$	0.7100	0.6801	0.6790	0.0930
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 0)$	0.1022	0.4159	0.2763	0.3664
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 0)$	0.0000	0.0425	0.0218	0.0538
$E(H_t \mid D_t^* = 1, D_{t-1} = 1)$	2.6325	2.6442	2.6401	0.0364
$E(H_t \mid D_t^* = 0, D_{t-1} = 1)$	2.6102	2.5971	2.5978	0.0255
$E(H_t \mid D_t^* = 1, D_{t-1} = 0)$	2.5860	2.5756	2.5828	0.0322
$E(H_t \mid D_t^* = 0, D_{t-1} = 0)$	2.5631	2.5736	2.5802	0.0218
$Pr(D_t^* = 1)$	0.5555	0.3315	0.2140	0.2591

B.4.2 General Approach

We combine $H_\tau = 1, 2$ and 3 and get the estimation results in Table B.9. The results are qualitatively similar to those in Table 2.11. One advantage for this setup over the previous one can be found from the ninth to the sixteenth row.

Table B.7: Estimation Results: Basic Approach for Males, Combining $H_\tau = 1, 2$ and 3

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 1)$	0.7443	0.8371	0.8421	0.1360
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 1)$	0.7443	0.6583	0.6585	0.1539
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 0)$	1.0000	0.4772	0.4102	0.3670
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 0)$	0.0817	0.0525	0.0431	0.0642
$E(H_t \mid D_t^* = 1, D_{t-1} = 1)$	2.6098	2.6316	2.6255	0.0461
$E(H_t \mid D_t^* = 0, D_{t-1} = 1)$	2.6098	2.5871	2.5892	0.0331
$E(H_t \mid D_t^* = 1, D_{t-1} = 0)$	2.5371	2.5454	2.5486	0.0353
$E(H_t \mid D_t^* = 0, D_{t-1} = 0)$	2.5695	2.5615	2.5671	0.0228
$Pr(D_t^* = 1)$	0.1309	0.3049	0.1944	0.2372

$E\left[H_t \mid D_t^*, D_{t-1}^*, d_{t-1}, h_{t-1} = 2.5\right]$ differ more in this table than in Table 2.11. This tells us that the estimation results are more reliable than those in Table 2.11 because the eigenvalues have more variation and thus can be better identified.

Table B.8: Estimation Results: Basic Approach for Females, Combining $H_\tau = 1, 2$ and 3

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 1)$	0.8666	0.8656	0.8770	0.1219
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 1)$	0.6689	0.6587	0.6554	0.1106
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 0)$	0.1719	0.3879	0.2553	0.3472
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 0)$	0.0000	0.0341	0.0151	0.0457
$E(H_t \mid D_t^* = 1, D_{t-1} = 1)$	2.6724	2.6830	2.6775	0.0679
$E(H_t \mid D_t^* = 0, D_{t-1} = 1)$	2.6066	2.5981	2.5983	0.0414
$E(H_t \mid D_t^* = 1, D_{t-1} = 0)$	2.6098	2.6158	2.6065	0.0603
$E(H_t \mid D_t^* = 0, D_{t-1} = 0)$	2.5890	2.5810	2.5912	0.0358
$Pr(D_t^* = 1)$	0.3589	0.3353	0.2118	0.2769

Table B.9: Estimation Results: General Approach, Combining $H_t = 1, 2$ and 3

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 1)$	0.8225	0.7621	0.7993	0.1658
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 1)$	0.4260	0.5900	0.5797	0.2153
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 1)$	0.2273	0.3488	0.3258	0.2302
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 1)$	0.1891	0.3071	0.2650	0.2502
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 0)$	0.6449	0.5098	0.5134	0.2218
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 0)$	0.7814	0.5337	0.5367	0.2471
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 0)$	0.2164	0.2630	0.2301	0.1908
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0)$	0.1256	0.0996	0.0471	0.1293
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.6096	2.5165	2.5897	3.6577
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	1.8385	3.2092	2.5267	23.3473
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.5168	2.7619	2.5317	1.0847
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	2.2861	1.7174	2.4180	6.7044
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	3.1835	3.1515	2.6905	6.7574
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	2.2231	2.1712	2.4967	2.6271
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	2.6110	2.6396	2.5865	1.2866
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	2.5846	2.1804	2.5810	13.3059
$Pr(D_t^* = 1)$	0.1926	0.2478	0.12093	0.1499

Table B.10: Estimation Results: General Approach of Males, Combining $H_t = 1, 2$ and 3

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 1)$	0.5158	0.7392	0.7575	0.1794
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 1)$	0.2811	0.5781	0.5803	0.2390
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 1)$	0.0000	0.3388	0.3299	0.2296
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 1)$	0.0000	0.3164	0.2501	0.2640
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 0)$	0.1608	0.3865	0.3419	0.2437
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 0)$	0.4849	0.4891	0.4920	0.2221
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 0)$	0.1608	0.2037	0.1507	0.1610
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0)$	0.0402	0.0853	0.0628	0.0922
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.5799	2.0940	2.5413	8.6562
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	2.5240	2.3785	2.5000	2.6628
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.3276	2.9407	2.5444	8.8653
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	2.3276	2.1845	2.4522	1.7809
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	2.7009	2.7169	2.5732	0.7657
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	2.4693	2.4031	2.4759	0.7829
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	2.7009	3.4648	2.5697	27.5230
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	2.5623	2.6390	2.5734	1.9286
$Pr(D_t^* = 1)$	0.2869	0.2931	0.2395	0.1915

Table B.11: Estimation Results: General Approach of Females, Combining $H_t = 1, 2$ and 3

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 1)$	0.9337	0.8286	0.8713	0.1513
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 1)$	0.7523	0.6528	0.6596	0.1934
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 1)$	0.3020	0.3853	0.4030	0.2285
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 1)$	0.2378	0.4072	0.4259	0.2153
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 0)$	1.0000	0.5619	0.5724	0.2502
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 0)$	0.1354	0.5443	0.5684	0.2747
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 0)$	0.1162	0.2627	0.1992	0.2151
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0)$	0.0000	0.0774	0.0282	0.1167
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.6982	2.6954	2.6696	2.9087
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	-0.0859	2.5871	2.5063	0.3876
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.5000	4.6111	2.5132	34.1491
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	2.4168	2.0740	2.5000	3.4499
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	2.7177	3.5561	2.7918	6.5064
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	2.5971	1.7223	2.5308	4.7850
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	3.5107	2.8225	2.5991	5.4624
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	1.2059	2.5537	2.5947	6.8411
$Pr(D_t^* = 1)$	0.9377	0.2428	0.1729	0.2007

Table B.12: Testing Validity of Conditional Independence – $H_t = 1, 2, 3$ Case

Null hypothesis H_0	95% confidence interval of $L(i)$
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 0)$	[0.0000,0.9659]
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 0)$	[0.4287,0.8532]

Table B.13: Testing Validity of Conditional Independence – $H_t = 1, 2, 3$ Case

Null hypothesis H_0	95% confidence interval of $L_3(d_1, d_2)$
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^*, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^*, D_{t-1} = 0)$	[0.0626,1.1417]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1}) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1})$	[0.0000,0.7888]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1} = 0)$	[0.0091, 0.8332]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1} = 1)$	[0.0000, 0.6127]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1} = 0)$	[0.0107, 0.8966]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1} = 0) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1} = 1)$	[0.0186, 0.8169]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1} = 0) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1} = 0)$	[0.0000, 0.6427]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1} = 0)$	[0.0073, 0.8706]

Curriculum Vitae

Ryan Bush was born in Columbus, Ohio on July 31, 1987. He received a B.A. degree in Mathematics from The University of North Carolina at Chapel Hill in 2010. He entered the Ph.D. program in Economics at Johns Hopkins University in 2010. He will start his work at the Federal Reserve Bank of New York in the summer of 2016.